

# Sādhana

## Academy Proceedings in Engineering Sciences

(Volumes 1-6 published as Proceedings of the Indian Academy of Sciences,  
Engineering Sciences)

### Editor

N Viswanadham

*Indian Institute of Science, Bangalore*

### Associate Editor

G Prathap

*National Aeronautical Laboratory, Bangalore*

### Editorial Board

D Banerjee, *Defence Metallurgical Research Laboratory, Hyderabad*  
I Chopra, *University of Maryland, College Park, MD*  
B L Deekshatulu, *National Remote Sensing Agency, Hyderabad*  
S C Dutta Roy, *Indian Institute of Technology, New Delhi*  
Y Jaluria, *Rutgers University, New Brunswick, NJ*  
K Kasturirangan, *ISRO Satellite Centre, Bangalore*  
B D Kulkarni, *National Chemical Laboratory, Pune*  
R Narayana Iyengar, *Indian Institute of Science, Bangalore*  
M A Pai, *University of Illinois, Urbana-Champaign, IL*  
P Ramachandra Rao, *National Metallurgical Laboratory, Jamshedpur*  
V U Reddy, *Indian Institute of Science, Bangalore*  
R K Shyamasundar, *Tata Institute of Fundamental Research, Bombay*  
J Srinivasan, *Indian Institute of Science, Bangalore*  
S P Sukhatme, *Indian Institute of Technology, Bombay*  
C E Veni Madhavan, *Indian Institute of Science, Bangalore*  
M Vidyasagar, *Centre for Artificial Intelligence and Robotics, Bangalore*

### Editor of Publications of the Academy

G Srinivasan

*Raman Research Institute, Bangalore*

---

### Subscription Rates

(Effective from 1989)

All countries except India (Price includes AIR MAIL charges)	1 year	3 years	5 years
	US\$ 75	\$ 200	\$ 300
India	1 year	10 years (for 1989-1998)	
	Rs. 75	Rs. 400	

All correspondence regarding subscription should be addressed to **The Circulation Department** of the Academy.

---

### Editorial Office

Indian Academy of Sciences, C V Raman Avenue  
P.B. No. 8005, Sadashivanagar  
Bangalore 560 080, India

Telephone: (080) 3342546  
Telex: 0845-2178 ACAD IN  
Telefax: 91-80-3346094

© 1994 by the Indian Academy of Sciences. All rights reserved.

"Notes on the preparation of papers" are printed in the last issue of every volume.

# **Sādhana**

(Academy Proceedings in Engineering Sciences)

**Volume 19**

**1994**

Published by the Indian Academy of Sciences  
Bangalore 560 080



# Sādhana

## Academy Proceedings in Engineering Sciences

(Volumes 1-6 published as Proceedings of the Indian Academy of Sciences,  
Engineering Sciences)

### Editor

N Viswanadham

*Indian Institute of Science, Bangalore*

### Associate Editor

G Prathap

*National Aeronautical Laboratory, Bangalore*

### Editorial Board

D Banerjee, *Defence Metallurgical Research Laboratory, Hyderabad*  
I Chopra, *University of Maryland, College Park, MD*  
B L Deekshatulu, *National Remote Sensing Agency, Hyderabad*  
S C Dutta Roy, *Indian Institute of Technology, New Delhi*  
Y Jaluria, *Rutgers University, New Brunswick, NJ*  
K Kasturirangan, *ISRO Satellite Centre, Bangalore*  
B D Kulkarni, *National Chemical Laboratory, Pune*  
R Narayana Iyengar, *Indian Institute of Science, Bangalore*  
M A Pai, *University of Illinois, Urbana-Champaign, IL*  
P Ramachandra Rao, *National Metallurgical Laboratory, Jamshedpur*  
V U Reddy, *Indian Institute of Science, Bangalore*  
R K Shyamasundar, *Tata Institute of Fundamental Research, Bombay*  
J Srinivasan, *Indian Institute of Science, Bangalore*  
S P Sukhatme, *Indian Institute of Technology, Bombay*  
C E Veni Madhavan, *Indian Institute of Science, Bangalore*  
M Vidyasagar, *Centre for Artificial Intelligence and Robotics, Bangalore*

### Editor of Publications of the Academy

G Srinivasan

*Raman Research Institute, Bangalore*

---

#### Subscription Rates

(Effective from 1989)

	1 year	3 years	5 years
	US\$ 75	\$ 200	\$ 300
All countries except India (Price includes AIR MAIL charges)			
India	1 year Rs. 75	10 years (for 1989-1998) Rs. 400	

All correspondence regarding subscription should be addressed to **The Circulation Department** of the Academy.

---

#### Editorial Office

Indian Academy of Sciences, C V Raman Avenue  
P.B. No. 8005, Sadashivanagar  
Bangalore 560 080, India

Telephone: (080) 3342546  
Telex: 0845-2178 ACAD IN  
Telefax: 91-80-3346094

© 1994 by the Indian Academy of Sciences. All rights reserved.

"Notes on the preparation of papers" are printed in the last issue of every volume.

## Academy Proceedings in Engineering Sciences

Volume 19, 1994

## CONTENTS

Artificial neural networks for pattern recognition	<i>B Yegnanarayana</i>	189-238
Convergence of higher-order two-state neural networks with modified updating	<i>M Vidyasagar</i>	239-255
Script recognition	<i>P V S Rao</i>	257-270
Solid finite elements through three decades	<i>D N Venkatesh and U Shrinivasa</i>	271-287
Development of robust finite elements for general purpose structural analysis	<i>G Prathap, B P Naganarayana and B R Somashekar</i>	289-309
Distortion, degeneracy and rezoning in finite elements - A survey	<i>Raviprakash R Salagame and A D Belegundu</i>	311-335
Magneto-visco-elastic surface waves in stressed conducting media	<i>Samar Chandra Das, D P Acharya and P R Sengupta</i>	337-346
A tutorial survey of reinforcement learning	<i>S Sathiya Keerthi and B Ravindran</i>	851-889
Large dynamic scheduling in manufacturing systems using Brownian approximations	<i>K Ravikumar and Y Narahari</i>	891-939
Design of software for safety critical systems	<i>R K Shyamasundar</i>	941-970
A survey of Indian logic from the point of view of computer science	<i>V V S Sarma</i>	971-983
Rudiments of complexity theory for scientists and engineers	<i>V Vinay</i>	985-994
Managing interprocessor delays in distributed recursive algorithms	<i>V S Borkar and V V Phansalkar</i>	995-1003

Traffic engineering in ATM networks: Current trends and future issues <i>A Gravey, G Hebutterne, R R Mazumdar and C Rosenberg</i>	1005–1025
PVU and wave-particle splitting schemes for Euler equations of gas dynamics <i>S M Deshpande, N Balakrishnan and S V Raghurama Rao</i>	1027–1054

Sonochemical reaction engineering <i>K S Gandhi and R Kumar</i>	1055–1076
--	-----------

### **Special Issue on Advances in Aerospace Engineering**

Foreword <i>R Narasimha and H S Mukunda</i>	347–348
Dynamical characteristics of wave-excited channel flow <i>S Selvarajan and V Vasanta Ram</i>	349–360
Detailed study of complex flow fields of aerodynamical configurations by using numerical methods <i>A Das</i>	361–399
On the boundary-layer control through momentum injection: Studies with applications <i>V J Modi and T Yokomizo</i>	401–426
Design and analysis trends of helicopter rotor systems <i>Inderjit Chopra</i>	427–466
Computational aircraft dynamics and loads <i>K Appa and J Argyris</i>	467–485
Recent progress in dynamics and aeroelasticity <i>A R Upadhyaya and Keshab Panda</i>	487–507
Forced vibration and low-velocity impact of laminated composite plates <i>Asghar Nosier, Rakesh K Kapania and J N Reddy</i>	509–541

### **Special Issue on Artificial Intelligence and Expert Systems**

Foreword <i>V V S Sarma and B Yegnanarayana</i>	1–3
Understanding design: Artificial intelligence as an explanatory paradigm <i>Subrata Dasgupta</i>	5–21
Fault diagnosis of machines <i>H N Mahabala, A T Arun Kumar, R R Kurup and G Ravi Prakash</i>	23–50
A fuzzy expert system approach using multiple experts for dynamic follow-up of endemic diseases <i>Apurba Banerjee, Arun Kumar Majumder and Anupam Basu</i>	51–73

Legal counselling system	<i>M Shashi, K V S V N Raju and A Lakshminath</i>	75-92
A prototype expert system for interpretation of remote sensing image data	<i>L Chandra Sekhara Sarma and V V S Sarma</i>	93-111
Paninian framework and its application to Anusaraka	<i>Akshar Bharati, Vineet Chaitanya and Rajeev Sangal</i>	113-127
Reducing barriers of communication across Indian languages: An AI and ES approach to mass media	<i>Shreesh Chaudhary</i>	129-146
Significance of knowledge sources for a text-to-speech system for Indian languages	<i>B Yegnanarayana, S Rajendran, V R Ramachandran and A S Madhukumar</i>	147-169
Planning in bridge with thematic action	<i>Deepak Khemani</i>	171-188
<b>Special Issue on Computation Heat Transfer</b>		
Foreword	<i>Y Jaluria and J Srinivasan</i>	615-617
The differentially heated cavity	<i>S Paolucci</i>	619-647
Flow transitions and pattern selection of the Rayleigh-Benard problem in rectangular enclosures	<i>D Mukutmoni and K T Yang</i>	649-670
Buoyant plane plumes from heated horizontal confined wires and cylinders	<i>G Lauriat and G Desrayaud</i>	671-703
Interaction between a buoyancy-driven flow and an array of annular cavities	<i>M Molki and M Faghri</i>	705-721
Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces	<i>C J Hoogendoorn, C L Koster and J A Wieringa</i>	723-749
Natural convection of a radiating fluid in a square enclosure with perfectly conducting end walls	<i>A Yucel, S Acharya and M L Williams</i>	751-764
Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems	<i>Ram V Mohan and Kumar K Tamma</i>	765-783
Finite element analysis of internal flows with heat transfer	<i>M Srinivas, M S Ravisankar, K N Seetharamu and P A Aswathanarayana</i>	785-816

Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials	<i>S Gopalakrishna and Y Jaluria</i>	817-832
A novel enthalpy formulation for multidimensional solidification and melting of a pure substance	<i>A W Date</i>	833-850
<b>Special Issue on Disaster Mitigation</b>		
Managing disasters precipitated by natural hazards	<i>V K Gaur</i>	543-550
Tropical cyclone hazards and warning and disaster mitigation systems in India	<i>Porathur V Joseph</i>	551-566
Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories	<i>U C Mohanty</i>	567-582
On the prediction of storm surges	<i>P K Das</i>	583-595
Origin, incidence and impact of droughts over India and remedial measures for their mitigation	<i>D A Mooley</i>	597-608
Earthquake hazard in Indian region	<i>K N Khattri</i>	609-614
<b>Subject Index</b>		1077-1083
<b>Author Index</b>		1084-1086

## Notes on the preparation of papers

Authors wishing to have papers published in *Sādhana* should send them to

The Editor, *Sādhana*,  
Indian Academy of Sciences, C V Raman  
Avenue, Post Box No. 8005, Sadashivanagar,  
Bangalore 560 080, India

*Four copies of the paper must be submitted.*

Submission of the script will be held to imply that it has not been previously published and is not under consideration for publication elsewhere; and further that, if accepted, it will not be published elsewhere.

**Aims and scope:** The journal will cover all branches of Engineering Science including mechanics (fluid, solid, thermal), computer science, electronics, energy, aerospace technology, materials science, nuclear engineering, systems analysis, alternative technologies etc. The editorial board wishes particularly to encourage papers (i) likely to be of interest to more than one professional group, either because the work is fundamental or because it reflects the best in current technology; (ii) summarising, at a sufficiently technical level, work done on special projects of interest to engineering scientists; (iii) offering extensive critical reviews, especially of subjects of interest in the country. Correspondingly, the board considers the publication of highly specialized work more appropriate to the many professional journals already being published. The criteria for acceptance of papers in the journal are therefore breadth of interest, irrespective of whether the paper reports research or development, theory or experiment, original work or review, and quality of work reported and its presentation (text, diagrams etc.). Papers on developments in Indian technology must pay special attention to the problems peculiar to the country in such projects.

**Typescript:** All parts of the script must be typed double spaced on one side of white bond paper of A4 (292 × 210 mm) or quarto (280 × 215 mm) size. A 4 cm margin must be provided for insertion of printers instructions.

**Title page** must contain (i) the title (which must be short and contain words useful for indexing), (ii) the initial(s) and names(s) of the author(s) and the name and address of the institution where the work was done, and (iii) a brief running title of not more than 50 letter spaces.

**Abstract** must be informative and not just indicative, and must contain the significant results reported in the paper.

**Keywords**, not more than about six in number, must be provided for indexing and information retrieval.

**The text** must be divided into sections, generally starting with 'Introduction' and ending with 'Conclusions'. The main sections should be numbered 1, 2 etc., sub-sections 1.1, 1.2 etc., and further subsections (if necessary) 1.1a, 1.1b etc.

**Tables** must be numbered consecutively in arabic numerals in the order of occurrence in the text; they should be self-contained and have a descriptive title.

**Figures**, including photographs which must be glossy prints, should be numbered consecutively in arabic numerals in the order of occurrence in the text. Line drawings must be in India ink on good quality tracing paper or Bristol board, preferably of the same size as the text paper. Lines should be sufficiently thick (axes about 0.3 mm, curves about 0.6 mm). All letterings on the illustrations must be done in pencil only; they will be inked at the editorial office. Figure captions must be typed on a separate sheet.

**A list of symbols** must be provided, with each symbol identified typographically (e.g. Gr. alpha, script oh, Latin ell etc.) for the printer; this list will not appear in print. Authors may if necessary provide another list of symbols for the reader (to be printed).

**Mathematical material:** Equations must be written clearly, each on its own line, well away

from the text but punctuated to read with it. Complicated expressions should be avoided in the text; when absolutely necessary they should be displayed separately like equations. All equations must be numbered consecutively in arabic numerals with the number in parentheses near the right hand margin. Superscripts and subscripts (which must be indicated with  $\vee$  and  $\wedge$  respectively) must be kept as simple as possible; those of greater than second order (e.g. subscript on superscript on superscript) must be avoided. Similarly symbols or letters with accents or unusual marks placed above or below them must be avoided. 'Oh' and 'zero', K, k and kappa, 'ell' and one etc. must be clearly distinguished. Authors must indicate in pencil in the margin wherever special characters (Greek, German, script, scalar, vector, tensor, matrix etc.) are required. All other mathematical letter-symbols will be set in italic type. Vectors must be underlined by a wavy line and tensors by two wavy lines.

**Units** and associated symbols must invariably follow SI practice.

**References** should be cited in the text by author and year, not by number. If there are more than two authors, reference should be to the first author followed by *et al* in the text. References at the end of the paper should be listed alphabetically by authors' names, followed by initials, year of publication, title of the article, name of the journal (abbreviated

according to standard practice), volume number, and numbers of first and last pages. References to books should include: name(s) of author(s), initials, year of publication, title of chapter (if any), title of the book, edition if not the first, initials and name(s) of editor(s) (if any) [preceded by ed(s)], place of publication, publisher, and chapter or pages referred to. References to theses must include the year, the title of thesis, the degree for which submitted and the University.

**Footnotes** must be avoided if possible but when necessary should be numbered consecutively, and typed on a separate sheet.

**Appendices** should be labelled A, B etc., in order of appearance.

**Proofs:** Authors are requested to prepare the manuscript carefully before submitting it for publication to minimise corrections and alterations in the proof which increase publication costs. Page proofs sent to the authors together with the reprint order form must be returned to the editorial office within two days of their receipt.

**Reprints:** 50 reprints of each article will be supplied free of charge.

**Style manual** can be obtained free of cost from the editorial office.

# Artificial Intelligence and Expert Systems

## Foreword

Three decades of research in Artificial Intelligence (AI) and about fifteen years of experience with its first commercial products known variously as Expert Systems and Knowledge-Based Systems have led to wild speculations about the revolutionary nature of these technologies affecting a wide range of human activities from wars to businesses, and from medical diagnosis to legal applications etc. It appears that AI has matured from the days of solving puzzles and games, proving mathematical theorems or manipulation in the block world domain of robotics to the wider world of diverse science, engineering and business applications where artificial creatures (agents, actors etc.) attempt to solve problems in benign and hostile real worlds. This special issue of *Sādhana* brings out a collection of papers emphasizing these trends in a predominantly Indian context.

AI research has evolved with the aim of making computing machines smarter by developing programs capable of simulating and mimicking facets of human intelligence. Traditionally, such systems are built using a top-down design approach. Vast amounts of knowledge, from a few human experts, is acquired and stored to solve a narrow range of problems in a specific domain capturing competence along one dimension of a human being's intelligent behaviour. Simultaneously, psychologists and cognitive scientists are concerned with computer programs to provide models of human mental activities such as thinking, learning, perception, use of natural languages, speech and pattern recognition. It is not difficult to recognize the fact that intelligent action in a challenging environment cannot be obtained through the classical weak methods of AI involving heuristic search. Today, intelligent action is synonymous with adaptive behaviour of the agent and intelligent systems are built in a bottom-up manner.

The papers in this special issue deal with problem-solving in diverse disciplines incorporating some of the current trends in AI and expert system fields. The first paper by Dasgupta looks at the problem of designing an artifact as a very important intelligent activity in engineering and looks at the spectrum of design problems ranging from routine to creative designs. The main contribution of the paper is to provide an AI model for a design theory. The design process starts from the initial state of a set of requirements  $R$ , to the goal state of a final design  $D$ , satisfying the requirements,  $R$ , through a structured set of design options. This theory considerably enhances our understanding of engineering design and the author concludes that the same model can serve to enhance our understanding of the act of creative design or an invention.

The next two papers deal with the generic problem of diagnosis. Mahabala and his colleagues discuss different approaches for fault diagnosis of machines. They present four approaches based on fault-tree, rule, qualitative and quantitative models. They demonstrate the superiority of model-based approaches over the rule-based approach while dealing with practical situations, where it is difficult to obtain precise



knowledge of the domain in the form of rules obtained from an expert. Model-based approaches can take care of situations not envisaged *a priori*, including minor variations in design. The authors recommend a multi-model diagnostic system which localizes the fault using a causal model and sees if the fault can be further localized by qualitative models and resorts lastly to analytical models, if warranted. In the paper on expert system applications to medicine, Banerjee, Majumder and Basu propose a system capable of long term monitoring of a chronic disease such as leprosy. Novel features of this system are the use of multiple cooperating experts, provision for correction of a mistaken diagnosis or incorrect treatment and incorporation of fuzzy production rules and fuzzy reasoning.

The next paper by Shashi, Raju and Lakshminath shows how expert system technology can be exploited for legal counselling and present a system that represents the relevant knowledge effectively. The system is capable of predicting the judgement in a case according to the available knowledge on the case. The authors introduce a frame-like knowledge structure called lattice with two-dimensional attributes for implementing the knowledge-base of the system. The authors illustrate their model of the judgement prediction system with a specific criminal case involving theft.

Sarma and Sarma demonstrate the application of expert system technology for interpreting remote sensing image data. They utilize the Dempster-Shafer theory of evidence to obtain a plausible interpretation by combining information from diverse sources such as collateral data, experiential knowledge, and pragmatics. They use the system for obtaining urban land cover and land use information.

The next three papers deal with language and speech. All these papers suggest exploitation of common features of Indian languages. The main focus in the paper of Bharathi and his colleagues is to use AI technology for developing a machine translation system between two Indian languages. They propose and use the Paninian framework for machine translation between free word order languages. Based on Paninian theory, they introduce the idea of *Anusaraka* or language accessor, which allows a reader who knows one language to access or follow a text in another language. This enables one to overcome the language barrier for communication in Indian languages. The paper by Shreesh Choudary highlights the real issues in communication through Indian languages especially in the context of mass media such as newspapers, radio and television. He identifies several problems faced by these media while projecting a news item to different regions of the country. The author feels that developments in AI can be used effectively to reduce the communication barriers in a multilingual society like ours. The paper by Yegnanarayana and his colleagues describes the development of a text-to-speech system for Hindi using several knowledge sources such as intonation, duration and coarticulation, besides knowledge of speech production at segmental level. They show that it is essential to acquire and incorporate these rules if the system is to produce intelligible speech from an unrestricted text. The paper discusses methodology to acquire these knowledge sources for Hindi, which can be adopted for other Indian languages as well.

The final paper in the issue addresses the planning problem in AI. Planning is knowledge intensive. Planning in an uncertain domain is rarely considered in contemporary AI systems. The paper by Khemani looks at the problem of planning in a dynamic and uncertain domain exemplified by the declarer's play in the game of contract bridge. The paper proposes a two-stage mechanism for planning. In the first stage, partial plans are suggested by a theme-based planner. The planner employs knowledge structures called thematic acts. In the second stage, a scheduler continues

the actions suggested by thematic acts into a coherent plan. The author feels that such a scheme of planning is particularly attractive for planning in complex domains where partial plans for newer goals have to be incorporated into the larger plans of an agent who may be continuously interacting with the world.

It took us more than a year to bring out this special issue due to our desire to address issues and topics which are new and relevant to the Indian context. The papers illustrate the range of possible applications of AI technology. We are grateful to the authors, who have put up with our numerous suggestions for revisions and our reviewers for their critical comments which significantly improved the presentation.

Finally, we thank the editors of *Sādhana* for giving us this opportunity and for putting up with the delay in bringing out this issue. As guest editors, we thoroughly enjoyed our interactions with the authors and the editors in this task.

February 1994

V V S SARMA and B YEGNANARAYANA  
Guest Editors



## Understanding design: Artificial intelligence as an explanatory paradigm

SUBRATA DASGUPTA

Department of Computation, University of Manchester Institute of Science and Technology, Manchester, UK

Present address: The Center for Advanced Computer Studies, University of South Western Louisiana, Lafayette, LA 70504–4330, USA

E-mail: dasgupta@cacs.usl.edu

**Abstract.** A substantial part of the intellectual content of what H A Simon called the ‘sciences of the artificial’ is contained in the activity we call *design*. A central aim of *design theory* is to construct testable, explanatory models of the design process that will serve to enhance our understanding of how artifacts are, or can be, designed. In this paper, we discuss how some of the basic concepts underlying the discipline of *artificial intelligence* (AI) can serve to provide an *explanatory paradigm* for understanding design. We present an AI-based model of the design process and describe some of the implications of this model for our understanding of design – including that aspect of it we call ‘invention’.

**Keywords.** Design; artificial intelligence; explanatory model of design; creativity; invention.

### 1. Design theory

Anyone who devises a course of action to change an existing state of affairs to a preferred one is involved in the act of design. As such, design is of central concern not only in traditional engineering – dealing with *material* artifacts such as structures, machines, circuits and production plants – but also in the generation of *symbolic* devices such as plans, organisations and computer programs. Indeed, it is this larger sense of the word ‘engineering’ that Herbert Simon had in mind when, in 1969, he coined the term ‘sciences of the artificial’ to designate all such disciplines that are concerned with the conception and production of useful artifacts (Simon 1981).

There has been a long-held notion that the sciences of the artificial (or, more conveniently, the *artificial sciences*) were simply *applications of the natural sciences*: that civil engineering, for example, is the application of mechanics, and mechanical engineering of mechanics and thermodynamics; or that electrical engineering is the application of electro-physics, and metallurgy of chemistry and solid state physics. The fact that the engineer and the researcher in the artificial sciences are concerned with the effecting of artifacts intended to serve some *purpose* and that purposiveness is totally at odds with the natural sciences hardly seemed relevant (according to

conventional wisdom) as far as the intellectual foundations of the artificial sciences were concerned.

Since the 1960s, several works have appeared which, in one way or another, have all been dedicated to the proposition that the world of the artificial contains its own logic which is related to but is quite distinct from the logic of the natural world (Jones & Thornley 1963; Pye 1964; Jones 1980; Cross 1984; Agüero & Dasgupta 1987; Brown & Chandrasekaran 1989; Coyne *et al* 1990; Dasgupta 1991). It has also come to be explicitly recognised that while there are many distinct artificial sciences – civil, mechanical, chemical and electrical engineering, metallurgy, aerospace technology, agriculture, computer science, organisation theory, economic and social planning, architecture etc. – there is one kind of intellectually nontrivial activity that is shared by all, viz., *design*. Furthermore, if we examine what the various kinds of designers have to say about their respective domains (be they bridges, machines, cities, software, administrative organisations or integrated circuit chips) we discover that the same *kinds* of things are being described regardless of the domain. The vocabulary may differ but the concepts are the same.

From such observations it has come to be realised that there is a significant component to all these domain-specific design processes that is essentially independent of what is being designed. That is, irrespective of whether we are designing chips, programming languages, computers, robots, airline reservation systems, bridges, cities or chemical plants, the processes of design have a strong *domain-independent* component.

The implication of this is considerable. For, it means that we can conceive of a discipline *the subject matter of which is the design process itself*. In recent years, this discipline has come to be known as *design theory* (Coyne *et al* 1990; Dasgupta 1991), and its scope or aim is essentially twofold:

- (a) to construct *explanatory models* of the design process – models that will serve to clarify, explain and enhance our understanding of the acts or processes whereby artifacts are, or can be, designed; and, consequently,
- (b) to establish foundations for the implementation of rational methods, tools and systems that may aid the activity of practical design.

These two objectives are mutually reinforcing in that each furthers the cause of the other: a better understanding of design as a cognitive process is likely to provide a sounder basis for inventing design methods and tools; conversely, the development and implementation of such methods and tools provide important data for constructing better explanatory models as well as for testing or evaluating such models. The two objectives also complement each other in that the first is concerned with the *description* of design viewed as an empirical cognitive process whereas the second relates to *prescribing* ways of doing design.

## 2. Artificial intelligence as an explanatory paradigm

In its ordinary sense, a *paradigm* is an example, a pattern or a model, as when we refer to the stored program computer as conceived in the 1940s as a paradigm for computer architecture, or as when Petroski recently referred to certain kinds of errors leading to engineering design failures, as 'paradigms for human error in design' (Petroski 1991).

This dictionary notion of paradigm was greatly enlarged in the 1960s by Kuhn

who in his (now classic) studies on the nature of scientific revolutions used this word in a very special way to advance an account of the origin and development of scientific disciplines (Kuhn 1962, 1970, 1977).

In essence, a Kuhnian paradigm is a network of generalised theories, metaphysical assumptions, metaphorical and heuristic models, methodological commitments, values and exemplars that are shared by, or are common to, a given scientific community. A paradigm provides the framework within which members of that community recognise and solve problems.

Kuhn's theory of paradigms and the role he ascribed to them in the development of scientific thought has been the subject of considerable discussion and criticism (Shapere 1964; Lakatos & Musgrave 1970; Laudan 1977, 1984; Suppe 1977; Lakatos 1978; Cohen 1985; Bohm & Peat 1987; Thagard 1990). Our concern here, however, is not with the nature of these arguments. In fact, we accept the essential substance of Kuhn's general thesis and wish to put it to a particular use. Our aim in this paper is to examine how the concepts underlying the discipline of *artificial intelligence* (AI) can serve as a Kuhnian paradigm for understanding the nature of the design process.

It is important to note that practically all research in the application of AI to the topic of design has been concerned with the prescriptive aspect of design theory – that is for automatising the design process (Mostow 1985; Brown & Chandrasekaran 1989; Chandrasekaran 1990; Coyne *et al* 1990; Gero 1991). This paper focuses on the role that AI plays or may play in the descriptive arena. More specifically, we shall be concerned here with the issue of how the concepts of AI can assist in the construction of *explanatory models* of the design act – including in the realm of the most creative level of design which we call invention. Thus, viewing AI as a Kuhnian paradigm for the exploration and understanding of design as a cognitive process makes it, in the context of this particular paper, not so much a technology as a theoretical handmaiden for cognitive science, much as mathematics serves as a servant for physics or for some aspects of engineering.

### 3. The metaphorical role of the artificial intelligence paradigm

The question that may obviously be raised at this stage is: how may AI be appropriate for this purpose? And why should AI be preferred as a basis for explanation (of design processes or any other phenomena demanding explanation) to some other paradigm? To answer these, we begin with the fact that explanations in many arenas of science – including cognitive science – frequently draw upon the use of *metaphors*; and that metaphors of a particular kind serve as *models*. Computational schemes of the type provided by AI are especially useful for the purpose of constructing such models. Let us elaborate on these points.

That metaphors are used as a means of understanding even in every day discourse is a commonplace idea. Indeed, Jaynes has made the point that understanding is *primarily* a matter of constructing metaphors whereby what we wish to understand (the *metaphrand*) is related to (or mapped onto) what we do understand or are familiar with (the *metaphier*) (Jaynes 1976).

What is less understood is that metaphors may play significant roles in scientific explanations. To take two celebrated examples, both Darwin and Lavoisier drew upon the use of metaphors to arrive at their respective conclusions about evolution and the chemistry of respiration (Gruber 1981; Holmes 1985).

As we have discussed elsewhere (Dasgupta 1993, 1994), the kind of metaphors

evoked by Lavoisier and Darwin are especially useful in that they serve to establish *analogical relationships* between metaphrand and metaphier. As a result, one can draw inferences or extract facts from the metaphier's domain and transfer them to that of the metaphrand where they can serve as sources of explanation. For example, in Lavoisier's case, the known chemistry of the burning of a candle (the metaphier) was exploited to suggest the unknown chemistry of respiration (the metaphrand) (Holmes 1985). In the case of Darwin, one of the metaphiers was artificial selection. This allowed him to draw upon facts pertaining to the hybridisation of plants and animals through breeding as a suggestive mechanism for how variations in species occur in nature (Gruber 1981).

Metaphors of these types, then, have instrumental or *heuristic* value because they can be used to *explain* as well as to evoke images. For this reason, it is more appropriate to call them *metaphorical models* (Dasgupta 1993, 1994). It is in this context that the language and concepts of AI are useful. For, if we are willing to accept that the act of design involves the *processing of symbolic structures* (see §4 and 5 below) then computation (in its most general sense) seems to provide the most appropriate tool for explaining the nature of such processes – since computation is, fundamentally, the discipline concerned with symbolic transformations and the processing of symbolic structures. Computation – and the particular form of computation that is the hallmark of AI – thus becomes a metaphorical model for explaining design.

It is important to emphasize, once more, the *heuristic* nature of such models. To take another instance from the history of science, the development of the kinetic theory of gases relied on 'seeing' gas molecules as hard, elastic and spherical – i.e., as microscopic billiard balls (Holton 1952). It is not *really* thought that gas molecules are billiard ball-like. Models are constructed and (tentatively) accepted *as if* they are true because it is useful or fruitful to do so. Viewing gas molecules *as if* they are hard, elastic spherical entities paved the way for classical mechanics to be applied in order to explain the known behaviour of gases.

Correspondingly, it does not have to be that computational models must capture the reality of the cognitive act of design in a 'truthful' way. Rather, we desire that such models should be able to *represent* design processes in the sense that:

- (a) the known or documented phenomena surrounding design acts can be explained by the model in a consistent way;
- (b) using the operational power of the model one can provide plausible explanations of cognitive acts of design for which there are no documented accounts;
- (c) the model provides a better explanatory framework than any other known paradigm.

In other words, if a computational account of design 'works', then, in the absence of a rival paradigm that 'works better', we should be willing to adopt, at least *tentatively*, the computation-based paradigm as an instrumental theory of the design process.

#### 4. A knowledge level model of the AI paradigm

At this time of writing there is, of course, something of a struggle between two schools of thought concerning the 'true' nature of the AI paradigm (Papert 1988). One is the *symbol processing model* which has its origins in the work begun in the 1960s by Simon, Newell and their collaborators (Newell *et al* 1960; Newell & Simon 1972, 1976;

Newell 1982) and the other is the *connectionist* model which, though having roots in the work of Pitts and McCulloch in the 1940s, assumed its modern form relatively recently (Papert 1988). Fortunately, this debate need not detain us here for *as far as design is concerned*, the dominant model is the symbolic version. Thus, in this paper at least, the AI paradigm is based on the symbol processing model.

To be more precise, we shall present a characterisation of the AI paradigm at what has come to be called the *knowledge level* of cognition. This term was actually coined, and a systematic treatment of its features first presented, by Newell (1982) although the knowledge level as an appropriate level at which cognitive processes could be described has long been tacitly recognised in the AI literature.

A system at the knowledge level will be referred to as an *agent*. The main entities with which an agent is concerned are *goals*, *actions* and *knowledge*. As Newell (1982) put it:

To treat a system at the knowledge level is to treat it as having some knowledge and some goals and believing it will do whatever is within its power to attain its goals insofar as its knowledge indicates.

In Newell's formulation, the connection between knowledge, goals and the choice of which action to take (in order to achieve the goals) is established by a behavioural principle which he termed

*The principle of rationality (PR)*: If an agent has knowledge that one of its actions will lead to one of its goals then the agent will select that action.

A problem with PR is that it tells us nothing about what the agent might do if it does not possess the requisite knowledge. Nor is it helpful in the situation where we observe an agent making a choice in response to a goal. Are we, for instance, to infer *abductively* that the agent possesses the requisite knowledge that that particular action will lead to the desired goal?<sup>1</sup>

Such a conclusion may be wholly unwarranted. An agent may possess *incomplete* or *partial* knowledge concerning the appropriate action to take in response to a goal. Alternatively, the *computational cost* of determining which action to select from a set of alternatives may be so *high* as to render such determination impractical. In other words, in addition to the rationality principle PR, an agent is governed by Simon's (1976, 1982)

*Principle of bounded rationality (PBR)*: Given a goal, an agent may not possess perfect or complete knowledge of, or be able to economically compute or access, the correct action (or sequence of actions) that will lead to the attainment of the goal.

The consequence of PBR for the theory of the knowledge level agent is that, given a goal, there is no guarantee that in selecting an action (or a sequence of actions) the goal will, in fact, be attained.

Ideally then, an agent's behaviour at the knowledge level is governed by PR. In reality, it is constrained by PBR. This means that any action(s) the agent chooses in

<sup>1</sup> Abduction is the rule of inference,

(IF A THEN B, B/A).

For a comprehensive discussion of abduction, see Thagard (1988)



order to attain a goal represents, in general a *hypothesis* (on the part of the agent) that the action(s) will lead to the goal.

An individual action *does* something. It has an *input* to it and it produces an *output*. In general, both input and the resulting output may be in the form of matter, energy or symbols. However, in the specific context of design, our concern is only with *symbol processing actions* in which the input and output are both symbol structures.

Symbols or structures composed out of symbols may, in general, be either *formal* (in that they stand for or represent mathematical sentences) or *physical* (in that they stand for or represent entities in some external universe – and so their ‘meaning’ are interpreted with reference to that universe). We shall use the term *general symbol structure* to refer to either formal or physical symbol structures.

We have noted above that the actions of interest here are symbol processing actions. In fact, actions may themselves be represented by symbol structures. More generally, *all goals, knowledge and actions pertaining to an agent are representable at the knowledge level by general symbol structures*.

Every action consumes some amount of time. While the actual duration of an action is unimportant here, it is to be recognised that an action has a *beginning* point and an *end* point in time; this means that an action may begin or end earlier or later than some other actions.

Actions may take place in sequence or in parallel. A *sequence* of actions  $a_1, a_2, \dots, a_n$ , where  $a_i$  ends before  $a_{i+1}$  begins ( $1 \leq i \leq n-1$ ) will, as a whole, have an input  $I$  which is the input to  $a_1$  and an output  $O$  which is the output of  $a_n$  such that the output of  $a_i$  is the input to  $a_{i+1}$ . Actions may also be conducted in *parallel* by an individual agent or a team of agents. It is assumed that parallel actions satisfy

*The principle of determinacy (PD):* If a set of actions  $a_1, \dots, a_n$  are conducted in parallel and if  $I$  is the input to this set then the output  $O$  will be identical to the output  $O'$  which would be produced if the same actions  $a_1, \dots, a_n$  were to be conducted in some *arbitrary* sequential order with the same input  $I$ .

In other words, according to PD, the input/output behaviour of a set of parallel actions is indistinguishable from the input/output behaviour of the same set of actions performed in *any* sequential order.

We shall refer to any sequential or parallel set of actions as a *structured set* of actions. Such a set will have one or more actions that are its *earliest* if no action outside this subset begins earlier than those within the subset.

Upto this point, actions have been linked with goals – that is, actions are assumed to be invoked in response to goals subject to the behavioural principles PR and PBR. However, it may also be possible for an action to be initiated *without* the stimulus of a goal. It may be initiated by virtue of an element or *token* in the agent’s knowledge body – in which case, such an action is not governed by PR or PBR. We shall, therefore, distinguish between *rational* actions (actions that are invoked in response to goals) and *nonrational actions* (those that are invoked in response to tokens in the agent’s knowledge body).

In summary, actions and the conditions of their invocation can be characterised as follows.

- [1] The input to an action is one or more symbol structures representing goals or knowledge tokens. If at least one of the inputs is a goal, the action is termed rational. Otherwise, it is nonrational.

- [2] The output of an action is one or more symbol structures representing either a knowledge token or a goal.
- [3] Every action entails the retrieval and application of tokens contained in the agent's knowledge body.
- [4] The choice of an action in response to a goal is governed by the principle of rationality (PR). That is, if an agent has knowledge (where such knowledge may be as weak as a belief) that one of its actions will lead to the goal being achieved, it will select that action.
- [5] Because of the bounded rationality principle (PBR), however, an action so chosen may not be the correct action or may not be economically computable by the agent.
- [6] Every action consumes time.
- [7] Actions may be performed sequentially or in parallel. A set of actions, some of which are sequential, others parallel, is said to be structured.
- [8] In a structured set of actions, its parallel subsets obey the principle of determinacy (PD).

Finally, the AI paradigm as a whole can be concisely described in the following terms:

#### DEFINITION 1

A *knowledge level process*  $P(KL)$  is a structured set of actions conducted by an agent (or, cooperatively, by a collection of agents) in response to a goal (or a conjunction of goals)  $G$  such that:

- (a) The input to  $P(KL)$  is a set of symbol structures at least one of which represents  $G$ .
- (b) The output of  $P(KL)$  is a set of symbol structures that represent goals or knowledge tokens where the latter includes, possibly, a *solution* to  $G$  – that is, tokens that represent a solution to, or achievement of,  $G$ .
- (c)  $P(KL)$  terminates when either (i) its output contains a solution to  $G$  or (ii) its output is such that no further action is (or can be) selected. *End Def*

Thus, the AI paradigm is defined here in the form of a symbol-transforming process. Such a process begins with a goal. The latter, subject to the principle of rationality, prompts an action (or a structured set of actions) involving the selection of tokens from the agent's knowledge body. The output produced by the action(s) may be a symbol structure which the agent believes is a (possibly partial) solution to the original goal.

However, because of bounds on the agent's rationality, the output may be a new (and more tractable) goal. The latter prompts one or more new actions to be performed and so the process continues. The process terminates when the original goal is achieved or when no further action can be performed by the agent.

### 5. Design as a knowledge-level process

One of the very real problems encountered by design theorists is the difficulty of defining the act of design in a form which, on the one hand, satisfies our intuitive idea of design and, on the other, permits useful and interesting inferences to be extracted from the definition. As we have discussed elsewhere (Dasgupta 1991), the many definitions advanced by theorists in the past have proved rather unsatisfactory

in these two collective aspects. Thus, rather than beginning with a definition, we may be forced to rely on our intuitive notion of design and examine its many characteristics in an empirical fashion. This was the approach we took, for example, in a previous work (Dasgupta 1991).

However, we believe that the knowledge level model of the cognitive agent as just described does provide the basis for the definition we seek – and herein lies the first benefit of the knowledge level AI model as a Kuhnian paradigm for design theory. Thus, we have:

## DEFINITION 2

A *design process* is a knowledge level process that satisfies the following properties:

- (a) The input to the process designates (or specifies or represents) a set of properties to be met by some artifact in some given universe. These properties are referred to as the *set of requirements*, *R*.
- (b) The output of the process designates (or represents) the artifact. This representation is referred to as the *design*, *D*.
- (c) The goal of the agent in conducting the process is to produce a representation or design (*D*) such that if an artifact is implemented according to *D* then it will satisfy the properties constituting *R*. This goal is referred to as the *design goal* and may be stated tersely as *D satisfies R*.
- (d) The agent has no knowledge of any design that satisfies *R*. *End Def.*

Let us consider, first, how this definition coheres with what we know empirically about the design process.

- (i) According to the above, design, being a knowledge-level process, is a structured set of actions that can be conducted by an individual agent or a team of agents. Thus, the definition recognises that design may be performed by a single designer or by a design team.<sup>2</sup>
- (ii) The actions performed do not lead to matter or energy to be transformed. They are *symbol processing* actions.<sup>3</sup>
- (iii) Furthermore, both the input and the output symbol structures designate entities in some given universe: the (input) requirements designate properties demanded of some artifact; the (output) design represents the artifact itself. The symbol structures are, then, *physical* symbol structures. The definition, thus, excludes purely formal symbol processing activities such as mathematics from its scope. This is intuitively satisfactory: we do not normally think of constructing theorems or proofs of theorems as designing.
- (iv) Because the output of a design process, as defined above, is a (physical) symbol structure, what it produced is a representation of the artifact, never the artifact itself. It is the representation that constitutes 'the design'. Thus, the definition allows us to distinguish between 'designing' and 'making' (Alexander 1964; Jones 1980; Dasgupta 1991). Obviously, the traditional craftsman of old also conceptua-

<sup>2</sup> For convenience, we shall talk simply in terms of *an agent* with the understanding that whatever is said applies equally to a team of agents.

<sup>3</sup> Of course, at some lower levels of abstraction (e.g. at the neuronal level) symbol processing actions will entail the transformation of matter and energy.

lised the form of the artifact he was creating. However, the essence of design is that it results in a *symbol structure* – that is the result is externalised and, consequently, communicable.

- (v) According to the definition, a design process is initiated only when the agent is posed with a set of requirements such that the agent is unaware of any other design or artifact satisfying the requirements. For a given set of requirements, if there already exists an artifact that satisfies it then there would be no need to design. Thus, 'newness' or change (in even the most modest of terms) is a condition for a design process to be initiated, according to the definition. This conforms to the observation that one designs in order to initiate change (Simon 1981; Dasgupta 1991).
- (vi) However, note that according to definition 1, 'newness' is always in the context of, or relative to, the agent's knowledge body. Empirically, this is entirely reasonable. For example, given a specification of requirements *R*, for an integrated circuit chip, the experienced engineer (i.e., one whose knowledge body contains many 'cases' – exemplars, in Kuhn's terms (Kuhn 1962, 1977) – of prior designs) may know of a chip or a chip design that satisfies *exactly* the requirements. In that case, there is no need to design the chip. The same problem given to a student or a neophyte engineer may lead to a design process being initiated simply because the latter has no knowledge of appropriate exemplars.

Most real design situations fall within these two extremes. For instance, the civil engineer is posed with requirements for a new bridge that include details of the required span, the local soil conditions, the topography and the expected loads. These specifications may be found to be similar but not identical to the characteristics of a particular design known to that engineer. In that case, a design process will be initiated which takes the known bridge design as the starting point. Thus, definition 2 excludes neither 'design from scratch' nor 'redesign'.

- (vii) It is well known that many design problems belong to the class of what Simon (1973) termed *ill-structured* problems. That is, the requirements are incomplete or imprecise or ambiguous or the space of potential solutions is unbounded.<sup>4</sup> In the rarer situations, design problems may be *well-structured* – that is, the requirements are stated in such a manner that one can immediately devise tests to determine whether or not a given design satisfies those requirements (Dasgupta 1991).

Along a different axis, the requirements may be such that the collective (or 'public') knowledge body of the relevant design community has no tokens that may provide the starting basis for a solution. In that case, the agent has to literally *invent a new artifactual form*. This situation corresponds to the most creative form of design, viz., invention. At the other extreme, the requirements may be such that the agent's knowledge body (or that of the relevant design community) has a very precise *archetypal form* or *schema* for the artifact. In that case, the design act may entail *instantiation* of the schema by fixing or setting some parameters to specific values. Brown & Chandrasekaran (1989) refer to this as *routine design*.

It will be noted that definition 2 allows for a range of design problems that fall within a space determined by both these axes (figure 1).

<sup>4</sup>Dasgupta (1991) gives many examples of ill-structured design problems

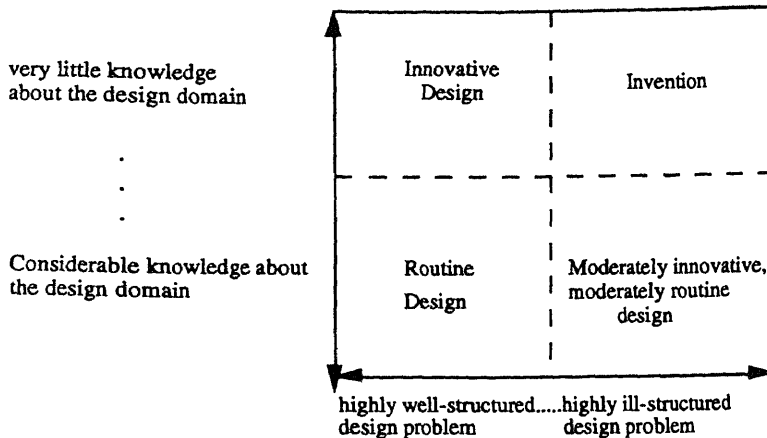


Figure 1. Space of design problems.

## 6. Implications of bounded rationality. I. Designs as satisficing solutions

Since the act of design is an instance of a knowledge level process, it is subject to the constraints of bounded rationality. This leads to several insights into the processes and nature of design. We consider some of these in this section and in the sections to follow.

One such insight is the distinction Simon made between 'optimal' and 'satisficing' designs. We have just seen that according to definition 2, design is a process entailing the construction of a representation of some artifact that meets the given requirements. It is understood that the solution sought is one that is the 'best' in some sense. In designing a computer system, for example, we seek to arrive at the best possible instruction set or memory management scheme or data path, as the case may be.

According to the principle of bounded rationality (PBR), however, even when all possible alternatives *are* known in advance, the cost of deriving an optimum or 'best possible' design may be prohibitively high. In spite of knowing that there exists an optimal solution to a design problem or even the actual procedure (that is, the set of knowledge level actions) that would yield the optimal, PBR tells us that the designer may not possess sufficient cognitive or computational resources to *actually determine* the optimum. Many of the optimisation problems encountered in design are what computer scientists call *intractable* in the sense that their solutions require processes of exponential time or (memory) space complexity (Dasgupta 1991).

So what does the designer *actually* do in the case of intractable problems? Or in the case of ill-structured design problems that are not amenable for formulation as optimisation problems?

One of Simon's major insights was that for most nontrivial design problems, levels of aspiration or *satisfactoriness* are established rather than criteria of optimality (Simon 1976, 1981). For instance, a bridge design is accepted if its estimated cost is 'below a certain amount'; a computer design project begins with the requirement that its peak performance must be 'twice that of its predecessor system' or have 'a better cost/performance ratio than that of its competitor'. If the design meets such criteria the problem is considered to have been solved. Simon named such solutions *satisficing*

solutions. Thus, in general, the design process attempts to satisfice rather than optimise.

## 7. Implications of bounded rationality: II. Two laws of design

There is yet another important consequence of PBR as far as design is concerned. It is that there is no guarantee that the design process conducted by an agent will, *in fact*, meet the design goal. A design  $D$  produced in response to the given requirements  $R$  embodies a *hypothesis* that  $D$  satisfies  $R$ . More formally, Dasgupta (1992) has recently proposed, and provided arguments in support of

*the hypothesis law*: A design process that reaches termination does so through one or more cycles of hypothesis creation, testing and modification.

While this law was derived directly from the knowledge level definition of design (Dasgupta 1992) and, in particular, from PBR, some form or another of this law has been widely recognised in the literature of design theory though couched mostly in terms of the concept of evolution. For example, Chandrasekaran's (1990) concept of a class of design methods which he called *propose-critique-modify* is, clearly, along the lines of the hypothesis law although he does not quite claim that his model constitutes a universal characteristic of the design process. Dasgupta has previously described in detail, with many examples from the domain of computer systems, the general concept of *design as an evolutionary process* and the idea of a design as constituting a theory or a hypothesis (Dasgupta 1989b, 1991). Finally, an earlier more informally stated suggestion, that designs signify hypotheses, is due to Petroski (1985).

Another ramification of the knowledge level model of the design process (and of bounded rationality) is captured by a law presented by Dasgupta (1992) called

*the impermanence law*: A design in any given state is never guaranteed to remain in that state.

Here, 'state' refers to the *state of belief* that may be held about the hypothesis that the design satisfies the requirements. Possible states are defined according to the following.

### DEFINITION 3

A design  $D$  produced in order to achieve a goal  $G$ : ' $D$  satisfies  $R$ ' for a given set of requirements  $R$  is said to be

- (i) **VALIDATED** when an agent produces a structured set of actions  $T$  (called a *test*) drawn from some knowledge body  $K$  that demonstrates that  $G$  has been achieved.
  - (ii) **REFUTED** when an agent produces a test  $T$  that demonstrates that  $G$  has *not* been achieved.
  - (iii) **TENTATIVE** when an agent can produce neither a test  $T_1$  that demonstrates that  $G$  has been achieved nor a test  $T_2$  that demonstrates that  $G$  has not been achieved.
- End Def.*

The structured set of actions – the tests  $T$ ,  $T_1$ ,  $T_2$  in the above definition – may take many forms. It may involve invoking some items from the agent's knowledge body, e.g. some previously published analysis or data; the construction of a mathematical

proof; a simulation experiment; or experiments constructed on a prototype. The outcome of the tests performed constitutes the *evidence*.

## 8. The non-monotonicity of design

As in the case of the hypothesis law, one can provide arguments in support of the impermanence law (Dasgupta 1992). The significance of the latter is considerable for it asserts that any evidence we summon in support of a claim about a design (that is, that it does or does not satisfy the requirements) is *itself conjectural*. That is, the reasoning underlying any claims we make about a design, like all empirical reasoning, is *non-monotonic* in nature (Reiter 1987). No matter how sure we may be at time  $t_1$  that the design is in the validated state (because, say, the evidence at time  $t_1$  happens to be compelling), there is no guarantee that this state of affairs will remain so at some (possibly much) later time  $t_2$  – when new contrary evidence may have come to light. For example, a new set of tests may falsify the earlier claim about the design being in the VALIDATED state; or we may realise that our earlier reasoning was faulty; or the assumptions upon which we had staked our claim may be discovered to be wrong. Anyone of these will result in the design being shifted to the REFUTED state.

The *practical* implications of the impermanence law – that is, of the non-monotonicity of designs – is also considerable when we consider the prescriptive side of design theory (refer §1) in which the concern is to propose effective design methods and tools. For, if the impermanence law is indeed universal then any design method we

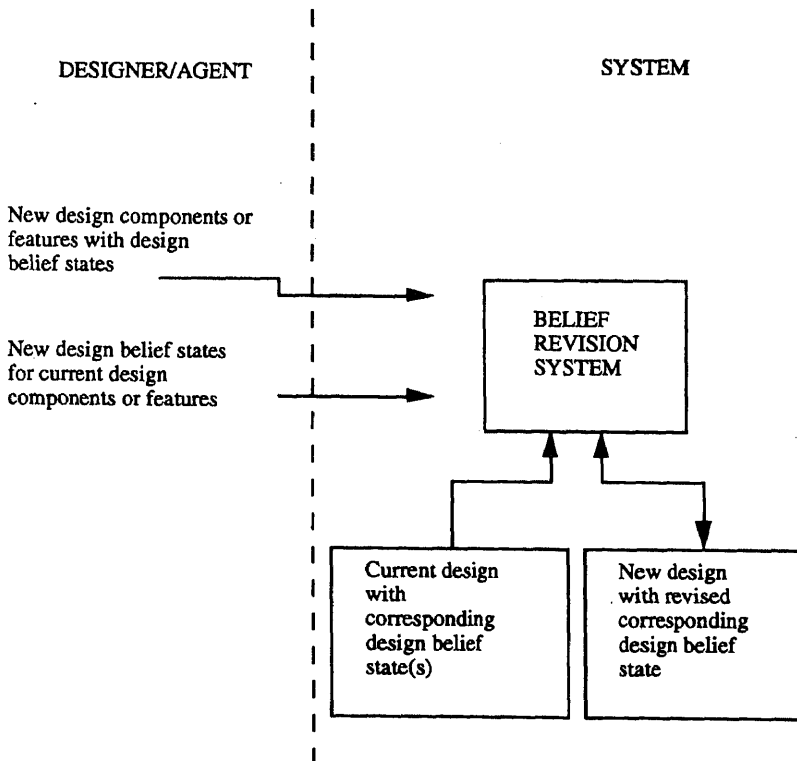


Figure 2. Architecture of a computer-aided belief revision system.

may propose (whether to be performed by 'cognitive' or 'computational' agents) must take into account the fact that the state of belief about the overall design *must be constantly revised* as and when new evidence is invoked or the design itself evolves or changes over time.

This fact – that the designer needs to constantly revise his or her claim about the design and maintain consistency amongst the belief states pertaining to different components of the design – was recognised explicitly and formed a central element of one practical design approach called the *theory of plausible designs* (TPD) developed by Dasgupta and his collaborators (Aguero & Dasgupta 1987; Hooton *et al* 1988; Dasgupta 1989b, 1991). In fact, this work demonstrated quite clearly that the need for belief state revision during design is so immediate that automating this aspect of the design process is virtually imperative.

In this regard, AI in its more computational *persona* provides additional benefits since the technique of what in AI is called *truth maintenance* (Doyle 1979; de Kleer 1986) can be applied. Patel & Dasgupta (1991) describe one such computer-aided system for belief revision, the overall architecture of which is outlined in figure 2.

## 9. Understanding invention

We noted in §6 that design problems may be mapped into some region of a space whose axes relate respectively to the 'structuredness' of the design problem and the amount of available knowledge about the relevant design domain (figure 1). The most intriguing region of this space concerns the *invention of original artifactual forms* which, as figure 1 suggests, is determined by situations where the design problem is highly ill-structured and virtually nothing is known about the nature or form of the artifact. Clearly, invention (or inventive design) is an aspect of the general problem of *creativity* in the artificial sciences and, thus, has many features in common with other kinds of creative acts, in particular, scientific discovery.

Consider, as a specific example, the invention by Wilkes (1951) of *microprogramming*. If creativity in the artificial sciences is strongly associated with the invention of new form then perhaps no better example can be found. For, the development of microprogramming led to an entirely new architecture for the control units of computers.

It is not our intention in this paper to present the technical details of microprogramming. For this, the reader may refer to any text on computer architecture (see, e.g., Dasgupta 1989a). However, the general history of the origins of microprogramming is well documented and is recounted here very briefly in order to illustrate the highly ill-structured nature of inventive design problems.

In the middle of 1949, the EDSAC computer, designed and built by Wilkes and his colleagues at the University Mathematical (later, Computer) Laboratory in Cambridge became the world's first fully operational 'stored program' computer. Soon after, Wilkes became preoccupied with the issue of *regularity* of computer designs. In particular, he was concerned with the fact that the organisation of EDSAC's control unit was irregular and *ad-hoc* (and, consequently, complex) in contrast to the highly regular organisation of EDSAC's memory unit.

What is interesting to note is that Wilkes *invented a problem*; moreover, it was a problem of a rather abstract kind, for it pertained to such qualities as 'regularity' and 'complex'. Wilkes problem was a *conceptual* problem (Laudan 1977; Dasgupta 1991) and such problems are inherently ill-structured. They are also of particular



interest in the context of creativity since the recognition of a conceptual problem by an individual is often motivated by philosophical or aesthetic viewpoints rather than strictly 'scientific' or 'technical' considerations.

In response to this particular problem, the principles of microprogramming were invented by Wilkes and first presented in a short paper (Wilkes 1951). Over the next two or three years, Wilkes and his colleagues developed the idea further and the first practical microprogrammed control unit was implemented in the EDSAC-2 which became operational in 1958 (Wilkes *et al* 1958).

Suppose now, we wish to investigate this process of invention; that is, we want to construct an explanation of how this cognitive act, performed by Wilkes, might have come about. How can we proceed?

Clearly, we cannot address the issue directly. What we might hope to do is to use the available historical evidence as recorded in the original papers, in subsequent retrospective accounts, in Wilkes's autobiographical memoirs and other sources (including personal communications and diaries) in order to construct a coherent structure of cognitive events which could serve as a plausible account of how Wilkes *might* have been led to his invention. The general idea, then, is to construct a *plausible model of creativity* which can explain this particular act of creativity in the realm of inventive design in a manner that is consistent with the historical evidence on hand.

We have recently completed a study of the invention of microprogramming using

---

#### A. The problem and observations

1. *The metaphrand*: The cognitive structure of creativity in the (natural and artificial) sciences.
2. *Relevant observations concerning the metaphrand*:
  - (a) A creative process involves changes in knowledge structures.
  - (b) Creativity involves the combination of known ideas or concepts with the resultant generation of novel ideas.
  - (c) The creative agent is purposeful and goal seeking.
  - (d) The creative process is protracted and evolving – and involves small changes of earlier ideas from moment to moment.
  - (e) Creative thinking entails searching for the 'right' ideas or concepts.
3. *Relevant observations concerning computation*:
  - (a) Computation entails the continuous modification of symbol structures.
  - (b) Computation begins with a goal and is directed, at all times, towards the attainment of the goal.
  - (c) 'Computations of a certain kind – 'knowledge level computations' – entail searching a space of possible and partial solutions with the aid of rules or heuristics to reduce the extent of search.

#### B. Formation of the metaphor

4. *The metaphor*: Scientific creativity as a cognitive process is like a knowledge-level computational process.
5. *The metaphier*: Knowledge-level computation.

#### C. Relevant knowledge about knowledge-level computation

6. The body of knowledge called (broadly) the 'Artificial Intelligence paradigm.'

#### D. Solution to the problem

7. A computation-based theory of scientific creativity.
- 

Figure 3. The structure of a computational metaphorical model of scientific creativity.

the knowledge level form of the AI paradigm as the basis for a metaphorical model of creativity. As discussed in §3, a metaphor entails the mapping of the unknown entity, the metaphrand, onto the known entity, the metaphier.

Figure 3 depicts the structure of our particular metaphorical model. As can be seen, relevant observations concerning the metaphrand and the metaphier (in this case, knowledge level computation) are listed in part A and are used to form the metaphor (part B). Once the metaphor is in place, the other relevant tokens of knowledge pertaining to the metaphier can be drawn upon to construct a 'computational theory of (scientific) creativity'.

It is not our intention, in this paper, to describe the details of this theory. We present it in great detail elsewhere (Dasgupta 1994). However, the general outline can certainly be given here. Basically, the 'computational theory of scientific creativity' is such that the process conducted by an agent leading to an original output can be described solely in terms of<sup>5</sup>

- (i) Symbol structures that represent goals, solutions and knowledge.
- (ii) Actions that operate upon symbol structures generating other symbol structures such that:
- (iii) Each symbol processing transformation is only a function of the agent's knowledge and the goal(s) to be achieved at that moment of time.

In other words our metaphorical model of creativity is that of the knowledge level agent described in §4. It is such that a creative process such as the one conducted by Wilkes can be described in the form of a knowledge-level process. The details of a plausible knowledge level process – plausible in that it is consistent with the historical and documented record – whereby Wilkes might have been led to the invention of microprogramming is described in Dasgupta (1994). Note that since the design process as previously described in §5 is itself a knowledge-level process, we arrive at the tentative conclusion that invention involves essentially the same kind of cognitive process as incurred in less creative acts of design. This is consistent with the conclusions reached by some others – both psychologists and computer scientists – who have investigated creativity (Newell *et al* 1962; Perkins 1981; Weisberg 1986; Langley *et al* 1987).

## 10. Conclusions

A substantial part of the intellectual content of the artificial sciences is contained in the activity we call design. A central aim of design theory is to construct testable, explanatory models of the design process that will serve to enhance our understanding of the processes whereby artifacts are or can be designed. The range of design problems include, at one extreme, routine design where the problem is very well-structured and there exists a large body of knowledge concerning the class of artifacts in question and, on the other, invention where the problem is highly conceptual, abstract and ill-structured and very little is known about the nature and form of the artifact.

---

<sup>5</sup>The criteria whereby an agent's output is deemed original must, of course, be quite independent of the theory of creativity. The latter attempts to explain how a creative process, i.e. a process the output of which is known to be original, may work. Elsewhere (Dasgupta 1993, 1994) we discuss in some detail the independent criteria whereby some cognitive act of discovery or invention may be judged to be original.

In this paper, we have discussed how some of the basic concepts underlying the discipline of artificial intelligence can serve to construct an explanatory Kuhnian paradigm within which the design process can be examined. The concept of a knowledge-level process provides such a paradigm. We have described here some of the implications of the knowledge-level model of design for our understanding of design and how the same model can serve to enhance our understanding of the act of invention.

## References

- Aguero U, Dasgupta S 1987 A plausibility driven approach to computer architecture design. *Commun. ACM* 30: 922-932
- Alexander C 1964 *Notes on the synthesis of form* (Cambridge, MA: Harvard University Press)
- Bohm D, Peat F D 1987 *Science, order and creativity* (New York: Bantam)
- Brown D C, Chandrasekaran B 1989 *Design problem solving* (London: Pitman)
- Chandrasekaran B 1990 Design problem solving: A task analysis. *AI Mag.* Winter: 59-71
- Cohen I B 1985 *Revolution in science* (Cambridge, MA: Harvard University Press)
- Coyne R D, Rosenman M A, Redford A D, Balachandran M, Gero J S 1990 *Knowledge based design systems* (Reading, MA: Addison-Wesley)
- Cross N (ed.) 1984 *Developments in design methodology* (New York: John Wiley & Sons)
- Dasgupta S 1989a *Computer architecture: A modern synthesis. Vol. 1. Foundations* (New York: John Wiley & Sons)
- Dasgupta S 1989b The structure of design processes. In *Advances in computers* (ed.) M C Yovits (New York: Academic Press) vol. 28, pp. 1-67
- Dasgupta S 1991 *Design theory and computer science* (Cambridge: University Press)
- Dasgupta S 1992 Two laws of design. *Intell. Syst. Eng.* 1 (Winter), 2: 146-156
- Dasgupta S 1993 Creativity, invention and the computational metaphor: Prologemenon to a case study. In *Creativity and artificial intelligence* (ed.) T Dartnall (Boston: Kluwer) (forthcoming)
- Dasgupta S 1994 *Creativity in invention and design* (Cambridge: University Press)
- de Kleer J 1986 An assumption based TMS. *Artif. Intell.* 20: 127-162
- Doyle J 1979 A truth maintenance system. *Artif. Intell.* 12: 231-272
- Gero J (ed.) 1991 *Artificial intelligence in design'91* (Oxford: Butterworth-Heinemann)
- Gruber H 1981 *Darwin on man: A psychological study of scientific creativity* 2nd edn (Chicago, IL: Univ. of Chicago Press)
- Holmes F L 1985 *Lavoisier and the chemistry of life* (Madison, WI: Univ. of Wisconsin Press)
- Holton G 1952 *Introduction to concepts and theories in physical science* (Reading, MA: Addison-Wesley)
- Hooton A, Aguero U, Dasgupta S 1988 An exercise in plausibility driven design. *Computer* 21: 7
- Jaynes J 1976 *The origin of consciousness in the breakdown of the bicameral mind* (Toronto: Univ. of Toronto Press)
- Jones C 1980 *Design methods: Seeds of human future* 2nd edn (New York: John Wiley and Sons)
- Jones C, Thornley D G (eds) 1963 *Conference on design methods* (Oxford, New York: Pergamon/Macmillan)
- Kuhn T S 1962 *The structure of scientific revolutions* (Chicago, IL: Univ. of Chicago Press)
- Kuhn T S 1970 Reflections on my critics. In *Criticism and the growth of knowledge* (eds) I Lakatos, A Musgrave (Cambridge: University Press) pp. 231-278
- Kuhn T S (ed.) 1977 Second thoughts on paradigms. In *The essential tension*. (Chicago, IL: Univ. of Chicago Press)
- Lakatos I 1978 *The methodology of scientific research programmes* (Cambridge: University Press)
- Lakatos L, Musgrave A (eds) 1970 *Criticism and the growth of knowledge* (Cambridge: University Press)
- Langley P et al 1987 *Scientific discovery* (Cambridge, MA: MIT Press)
- Laudan L 1977 *Progress and its problems* (Los Angeles: Univ. of California Press)

- Laudan L 1984 *Science and values* (Berkeley, CA: Univ. of California Press)
- Mostow J 1985 Towards better models of design processes. *AI Mag.* Spring: 44–57
- Newell A 1982 The knowledge level. *Artif. Intell.* 18: 87–127
- Newell A, Shaw J C, Simon H A 1960 Report on a general problem-solving program for a computer. *Information processing* (Paris: UNESCO) pp. 256–264
- Newell A, Shaw J C, Simon H A 1962 The processes of creative thinking. In *Contemporary approaches to creative thinking* (eds) H E Gruber, G Terrell, M Wertheimer (New York: Atherton) pp. 63–119, 933–951
- Newell A, Simon H A 1972 *Human problem solving* (Englewood-Cliffs, NJ: Prentice-Hall)
- Newell A, Simon H A 1976 Computer science as empirical inquiry: Symbols and search. *Commun. ACM* 19: 113–126
- Papert S 1988 One AI or many. *Daedalus* Winter; also *Proc. Am. Acad. Arts Sci.* 117: 11–14
- Patel S, Dasgupta S 1991 Automatic belief revision in a plausibility-driven design environment. *IEEE Trans. Syst. Man Cybern.* 21: 933–951
- Perkins D N 1981 *The mind's best work* (Cambridge, MA: Harvard Univ. Press)
- Petroski H 1985 *To engineer is human* (New York: St. Martin's Press)
- Petroski H 1991 Paradigms for human error in design. *Proc. 1991 NSF Design and Manufacturing Systems Conf.* Austin TX., Jan, pp. 1132–1146
- Pye D 1964 *The nature of design* (London/New York: Rheinhold/Studio Vista)
- Reiter R 1987 Nonmonotonic reasoning. *Annu. Rev. Comput. Sci.* 2: 147–186
- Shapere D 1964 The structure of scientific revolutions. *Philos. Rev.* 73: 383–394
- Simon H A 1973 The structure of ill structured problems. *Artif. Intell.* 4: 181–200
- Simon H A 1976 *Administrative behavior* 3rd edn (New York: The Free Press)
- Simon H A 1981 *The sciences of the artificial* 2nd edn (Cambridge, MA: MIT Press)
- Simon H A 1982 *Models of bounded rationality* (Cambridge, MA: MIT Press)
- Suppe F (ed.) 1977 The search for philosophic understanding of scientific theories. In *The structure of scientific theories* (Urbana, IL: Univ. of Illinois Press)
- Thagard P 1988 *Computational philosophy of science* (Cambridge, MA: MIT Press)
- Thagard P 1990 The conceptual structure of the chemical revolution. *Philos. Sci.* 57: 183–209
- Weisberg R W 1986 *Creativity: Genius and other myths* (New York: W H Freeman)
- Wilkes M V 1951 The best way to design an automatic calculating machine. *Rept. Manchester Univ. Comput. Inaugural Conf.*, Manchester, UK
- Wilkes M V, Renwick W, Wheeler D J 1958 The design of a control unit of an electronic digital computer. *Proc. Inst. Electr. Eng.* 105: B121



## Fault diagnosis of machines

H N MAHABALA<sup>†</sup>, A T ARUN KUMAR, R R KURUP and  
G RAVI PRAKASH

KBCS Laboratory, Department of Computer Science and Engineering,  
Indian Institute of Technology, Madras 600 036, India

E-mail: mahabala@iit.ernet.in

**Abstract.** This paper presents four major approaches for diagnosing machine faults. Given the description of a system to be diagnosed and the observations on the system when it works, the need for diagnosis arises when the observations are different from those expected. The objective of diagnosis is to identify the malfunctioning components in a systematic and efficient way. The four approaches discussed are based on fault-tree, rule, model, and qualitative model. Early diagnosis systems used fault-tree and rule-based approaches. These are efficient in situations where an expert is able to provide the knowledge in the form of associations between symptoms and faults. Model-based and qualitative model-based approaches overcome many of the deficiencies of the earlier approaches. Model-based approaches can take care of situations (faults) not envisaged *a priori*. Also, one can cater to minor variations in design using the same set of components and their interconnections. This paper discusses in each case, how the knowledge is represented and what diagnosis technique is to be adopted, and their relative advantages and disadvantages. Implementation of each method is also discussed.

**Keywords.** Rule-based diagnosis; fault-trees; model-based diagnosis; qualitative model-based diagnosis; power system diagnosis; multi-level qualitative reasoning.

### 1. Introduction

Diagnosis is one of the major application areas of knowledge-based systems today. If observations on a system in operation differ from the behaviour expected of the system, then the need for diagnosis arises. The goal of the diagnosis is to identify the malfunctioning components of the system.

There are many different approaches to diagnostic reasoning. Diagnostic fault-trees, rule-based reasoning, model-based reasoning, and qualitative model-based reasoning are the strategies used successfully in some narrow domains. Most of the early diagnostic systems used fault-trees (Williams *et al* 1983) and rule-based approaches (Vesonder *et al* 1983; Fukul & Kawakami 1986; Talukdar *et al* 1986).

Fault-trees provide a simple and efficient way to express the tests and conclusions

thereof needed to guide the diagnosis under various conditions. The diagnostic procedure traverses the tree starting from the root, applying the test at each node to decide the branch to be taken next, until it reaches the lead node (repair node).

In the rule-based approach, knowledge for diagnosis is captured in the form of IF-THEN rules. Rule-based systems are built by accumulating the experience of expert diagnosticians in the form of empirical associations between the symptoms of an abnormal system and the underlying faults.

Rule-based and fault-tree based approaches have some disadvantages in spite of their simplicity and efficiency. Both the systems are device-specific and must be reformulated even for a minor change in the device configuration. These systems are expensive to build and maintain. A small change in the device may require major restructuring of the tree or rules.

The model-based approach (de Kleer & Williams 1987; Struss 1988), the successor to the rule-based approach, attempts to overcome many of the limitations of the early systems. In the model-based approach, the device to be diagnosed is modelled and represented in terms of the structure and function of the individual components comprising the device. The correct behaviour of the device is inferred from the knowledge of the individual components and their interconnections. The model-based approach has the following advantages.

- The device description (function and structure) is represented explicitly;
- a domain-specific component library can be established;
- a device can be diagnosed by having a domain-independent diagnostic procedure which uses the model of the device derived from the library;
- it is possible to cover a large class of devices built out of the same set of components;
- diagnosis of new devices about which one does not have sufficient experience is possible;
- one can cover new symptoms and related faults.

A model-based approach requires a precise mathematical model of behaviour/functionality. But often, for diagnostic purposes, one can start with a description using trends and tendencies without resorting to precise quantification. The qualitative reasoning approach is based on such qualitative models and can often be used where models are not available or are too complex to deal with. The precision of the behavioural description in the quantitative models could be sacrificed by qualitative methods by retaining the crucial distinctions. Instead of continuous real-variables, each variable is described qualitatively using a small number of qualitative labels (e.g. +, - or 0). Quantitative differential equations are converted into qualitative differential equations called confluences (de Kleer & Brown 1984).

Using a single model for troubleshooting in all situations may not work. There may be need for multiple models of the same device with different simplifying and operating assumptions. The qualitative model-based approach has the following advantages, in addition to those of the model-based approach.

- Adopts an approach followed traditionally by practising engineers to describe behaviour and express malfunctions;
- derives qualitative behaviours which are adequate and efficient from precise mathematical models;
- since fuzzy models are very similar to qualitative models, one can use results from the fuzzy systems area.

## 2. Diagnosis using fault-tree

Fault-trees are a natural and efficient way to represent a hierarchical organization of tests and conclusions thereof needed for the diagnosis of industrial equipment. Diagnosis can be viewed as the task of hypothesizing the locality of a fault and then successively refining the hypotheses based on the results of a hierarchy of tests at each step. A scheme that captures such knowledge naturally and effectively is the fault-tree. However, one is not satisfied with just the localization of the faults in most cases. The fault needs to be repaired. Repair information can be stored in the fault-tree to supplement the diagnosis. In our implementation, called IITMDESS (Mahabala *et al* 1992), we go a step further and provide back-pointers to the tests at the repair nodes. These back-pointers help in ascertaining the effectiveness of the repair done and in determining the existence of multiple faults.

A fault-tree consists of a set of different types of nodes to represent the diagnostic knowledge. Node types correspond to different situations which arise while performing diagnosis. The current implementation uses six different types of nodes: Control, Repair, Text, Branch, Or, and Subtree. These different nodes are shown in figure 1.

### 2.1 Types of nodes

**Or node:** A symptom is usually associated with a set of possible causes (by cause, we mean a faulty component/subcomponent). In some domains, it is a practice to order these causes statically in a priority order. In other domains, one may order the causes dynamically depending upon the parameters like MTBF (mean-time between failures), MRRP (most recently replaced part) etc. An Or node captures and represents this information. It represents the likely causes as branches to subtrees. If information on static ordering is available, it is represented as weights on the branches to subtrees.

**Text node:** In the course of locating a fault, one may want to separate a subsystem (disassemble) before continuing further. For example, one may want to disconnect a

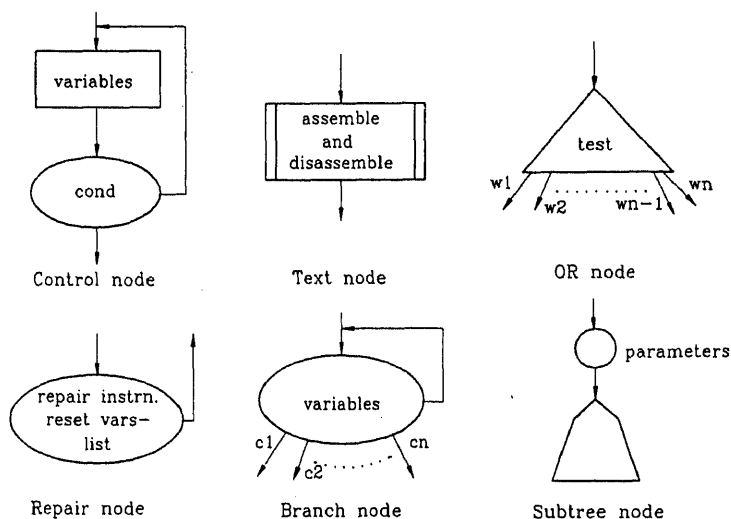


Figure 1. Different types of nodes of a fault tree.



power supply before diagnosing the equipment. Text node allows one to capture information that instructs the operator to disassemble sub-components on the way down and assemble them again on the return traverse.

*Control node:* This type of node decides whether there is a faulty condition of a certain category. It also specifies a list of variables which should be measured, and a conditional expression, which is evaluated to determine whether there is a fault of a specified type or not.

*Repair node:* When a diagnosis engine enters a repair node, one fault is identified. Repair nodes capture repair instructions, execution of which will eliminate the fault. Since repair may invalidate certain earlier measurements which need to be remeasured, one can also include a list of variables that should be reset after the repair action. One can associate a level of competence with a repair node which can be used to decide whether the current operator can handle the repair or not. The competence level is based both on training as well as access to tools needed. On completion of the repair, one has to go up in the fault-tree (rechecking certain measurements) to confirm the effectiveness of the repair, which is referred to as return traverse.

*Branch node:* During diagnosis, it is possible that multiple repairs are to be carried out to a single component. Branch nodes represent this knowledge. Branch nodes are similar to the control nodes, but they can have many child nodes. A branch node enables evaluation of multiple conditional expressions (one for each of its child nodes) and guides the diagnosis engine through one or more of its child nodes.

*Subtree node:* Complex equipment often have many components, subcomponents etc. Independent fault-trees can be developed for each subcomponent. A subtree node in the component level fault-tree enables diagnosis to enter an independent fault-tree corresponding to a subcomponent. This feature enables modular development of a fault-diagnosis system.

## 2.2 The inference mechanism

The inference mechanism consists of two steps. In the first step, we search for the cause of the abnormal behaviour of the machine in terms of repair needed. In the second step, after the repair has been done, a return traverse is used to check whether the problem has been completely eliminated. The basic reasoning cycle of the fault-tree based inference engine is described by the flow chart given in figure 2.

The processing starts from the root node traversing down dictated by the measurements and the type of nodes of the tree. When an Or node is entered, the system asks whether the corresponding symptom is observed. If so, the successor with the highest priority is selected for processing. Otherwise, the search backtracks to the parent node.

When a control node is entered, then the conditional expression associated with that node is evaluated using the values of the variables (which are measured if not already done before). If the condition is satisfied, then the successor node is explored.

When a branch node is reached, values are measured and the conditional expressions associated with the branch node are evaluated. If a variable has been measured earlier, it will not be measured again. The user is prompted whenever a measurement is

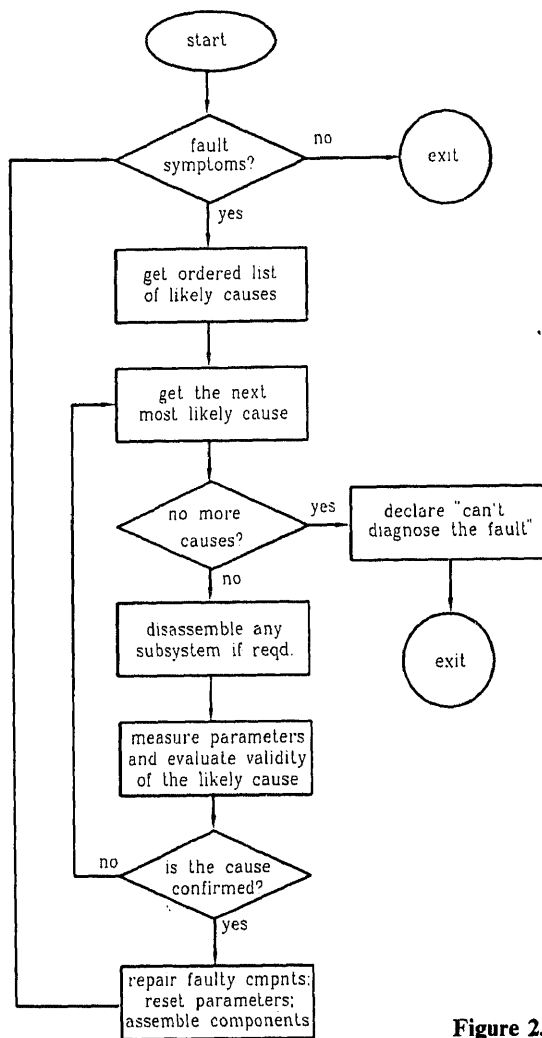


Figure 2. Reasoning cycle of the inference engine.

needed. If only one condition holds true, the corresponding branch is entered. If more than one condition holds true, its corresponding successors are processed in a left to right order or in the order of decreasing weights. If none of the conditions are satisfied, then the process backtracks to the parent node, deducing that the fault is elsewhere, and other paths are explored.

When a repair node is reached, the repair instructions are displayed by the system. After each repair, the parameters included in the list of resettable variables are reset and the search backtracks to the parent node. There is a provision to jump to a level higher than the parent, thereby making the checking of the elimination of fault more efficient.

While searching, if a subtree node is encountered, then during forward traverse the subsystem fault-tree is rolled in and the relevant parameters are passed to the subtree. The diagnosis continues from the root of the subtree. During return traverse, the subtree is rolled out and backtracking continues in the parent tree.

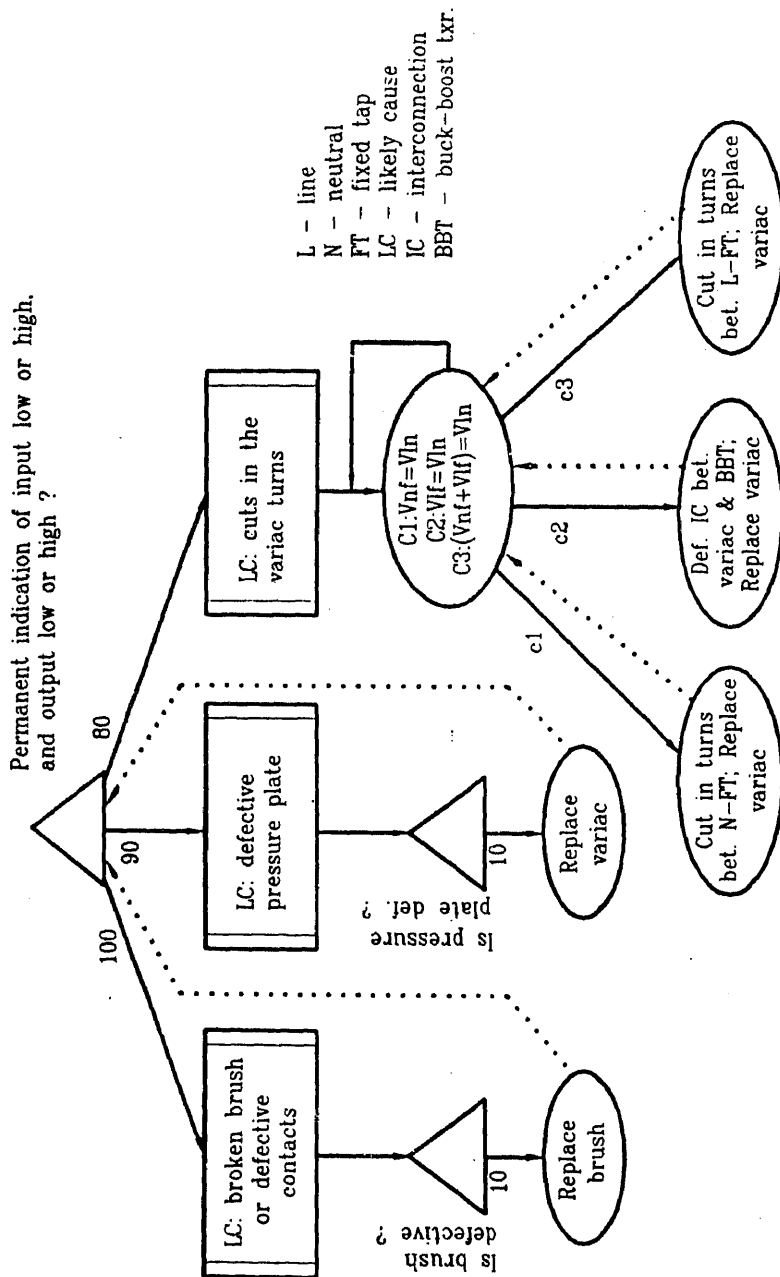


Figure 3. Part of a fault-tree of a servo-controlled voltage stabilizer.

### 2.3 Example

In this example, we illustrate the diagnosis of a particular brand of servo-controlled voltage stabilizer. A part of the fault-tree is shown in figure 3. Let us assume that there is a cut in the turns between the neutral and the fixed-tap of the variac. The diagnosis process starts from the root node and prompts the user whether there is any abnormality in the input-output panel indications. As the answer will be 'yes', the inference mechanism first checks whether the cause is the faulty brush. Having found that the cause is not the brush, the system verifies the cause is the faulty pressure-plate of the variac.

On obtaining a negative answer, the inference process verifies whether the cause is any cut in the variac taps. The system seeks values to the parameters,  $V_{ln}$ ,  $V_{nf}$  and  $V_{lf}$ . The three conditional expressions of the branch node are evaluated. Since we assumed that there is a cut in the turns between the neutral and the fixed tap of the variac, the system finds that ( $V_{nf} = V_{ln}$ ) and instructs the necessary repair. A part of the diagnosis session corresponding to this fault is shown in appendix A.

### 2.4 Special features of IITMDESS

The following are the special features of IITMDESS.

- (1) The system has a knowledge-acquisition module which provides an interactive environment for the expert to graphically edit the fault-trees;
- (2) the system has a facility to suspend the diagnosis session at any time and to resume at a later time. This feature helps the repairs, which take a long time, by relieving the computer for other uses;
- (3) the system makes use of the same fault-trees in providing a practically useful 'Training and Testing Facility' to the service engineers.

## 3. Rule-based diagnosis

Rule-based approach is an efficient and simple way of organizing the knowledge of a machine to be diagnosed in the form of rules. A rule takes the following forms:

```

IF <antecedent-1>
AND <antecedent-2>
.....
AND <antecedent-n>
THEN <consequent-1>
AND <consequent-2>
.....
AND <consequent-m>

```

Rules incorporate knowledge which relate symptoms with the underlying malfunction. Traditional rule-based systems have been built by accumulating the experience of expert diagnosticians in the form of rules. The association of symptoms with the underlying faults are based on the experience with the device to be diagnosed. For example, a fault in the motor-starting system would be represented as:

IF the engine does not turn over  
 AND the battery voltage is ok  
 AND the starter motor relay operates  
 AND the starter motor does not turn  
 THEN the starter motor is defective.

A rule can also conclude an intermediate condition (hypothesis). The rules are organised such that one or more intermediate hypotheses are deduced and then combined to produce the final diagnosis of the system.

The diagnosis starts with an observation (or problem) of the working system. Based on the problem, rules are selected and fired, gathering new observations until either a faulty component is diagnosed or the fault cannot be diagnosed with the available knowledge (rules). Generally, the user is queried for new observations/measurements as the diagnostic session progresses. When it identifies a faulty component, repair action or similar advice could be suggested. In the case of multiple faults, the diagnostic process can be continued after the repair until all faults are identified and repaired. One should adopt backward chaining (goal-driven) to make the system prompt the user. If the results of all tests are available at the start one can use forward chaining (data-driven).

A portion of the rule-base for PC diagnosis (for floppy disk problems) and a sample diagnosis session are given in the appendix B.

### 3.1 Implementation

A rule-based shell, IITMRULE, developed by the nodal centre on expert systems at our Institute has been used to implement the diagnostic system. IITMRULE is a powerful tool for building rule-based systems. It has a rule base editor and a compiler. It supports backward chaining as well as forward chaining of rules. Also, it has a database toolkit to access a database. The detailed description can be found in Mahabala & Ravikanth (1990).

## 4. Model-based diagnosis

Suppose one has an adequate model (usually mathematical) of a system. If there is a difference between the behaviour manifested by the system and that predicted by the model, the system is faulty and we need to identify the fault(s). The diagnostic task is to identify the faulty components which explain the discrepancy between the observed and the expected behaviour of the system. The general principles of the model-based diagnosis are shown in figure 4. The model-based diagnosis starts with:

- SD, a description of the system to be diagnosed given in terms of the model for the normal behaviour of its components and their interconnections, and
- OBS, a set of observations about the real system.

From the model and an initial set of values, the diagnostic system predicts the expected behaviour of the system. If all the components are functioning correctly, then  $SD \cup OBS \cup COMP$  is consistent, where COMP is the set of assumptions about the correctness of the components. If  $SD \cup OBS \cup COMP$  is inconsistent, then the need for diagnosis arises. The diagnoses  $\Delta$  is a set of components,  $\Delta \subseteq COMP$ , such that

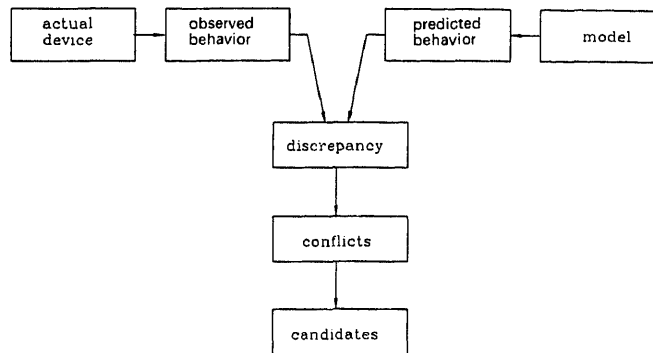


Figure 4. Principles of model-based diagnosis.

$SD \cup OBS \cup COMP^*$  is consistent, where  $COMP^*$  is obtained by replacing components in  $\Delta$  by their faulty models.

The diagnostic process is incremental. Diagnoses are determined by generating prediction about observable parameters of the device from subsets  $SD' \subseteq SD$  and  $COMP' \subseteq COMP$  and then checking these predictions for consistency with the observations  $OBS' \subseteq OBS$ . It is possible to gather evidence against correctness assumptions of the components by analysing the root-cause of discrepancies between predictions and observations. From this discrepancy, a set of components called conflict is generated. Conflict is a set,  $CONFL \subseteq COMP$ , such that some or all components in  $CONFL$  are faulty. That is, all the components in the conflict set cannot work correctly. More observations would produce more conflicts, and the set of all the conflicts is referred to as  $CONFLICTS$ . From the conflicts, sets of components called candidates (diagnoses) are generated. A candidate is a set  $CAND \subseteq COMP$  such that:  $\forall CONFL \in CONFLICTS [CAND \cap CONFL \neq \emptyset]$ . That is, a candidate must have at least one component from all the conflicts.

Usually, additional observations are necessary to isolate the set of components which are actually faulty. Each observation may produce new sets of conflicts and they are used to refine the candidates so far produced. The best next measurement is the one which will, on an average, lead to the discovery of the set of faulty components in a minimum number of measurements (de Kleer & Williams 1987).

An assumption-based truth maintenance system (ATMS) (de Kleer 1986) is used as a tool which records assumptions, inferences and their dependencies, observations, inconsistencies etc. The ATMS is a powerful tool for multiple context reasoning, and it is capable of working with multiple contexts simultaneously. This feature of the ATMS is used by the diagnosis system. Section 4.1 gives an overview of the basic ATMS, and §§ 4.2 and 4.3 discuss examples, algorithms, implementation and a practical application of the model-based diagnosis.

#### 4.1 Overview of the ATMS

Model-based reasoning systems, such as diagnosis systems, make inferences based on certain assumptions. These inferences have to be recorded with their dependencies for further reasoning. Also, inconsistencies have to be detected as and when they occur, to prevent their propagation. An ATMS provides a good framework to achieve

this, and forms one of the components of the overall problem solving system (de Kleer 1986b; Dressler & Farquhar 1990). Every inference made by the problem solver is communicated to the ATMS. The task of the ATMS is to keep track of these inferences and the associated assumptions which support those inferences.

ATMS works with a set of nodes and a set of justifications, where a node is an internal representation for a problem-solver datum, and a *justification* is the reason based on which a node has been justified. A justification is of the form  $a_1, a_2, \dots, a_n \Rightarrow c$ , where  $a_1, a_2, \dots, a_n$  are called *antecedents* and  $c$  is called *consequent*. A subset of antecedents are *assumptions*, whose truth or false value can get changed. For example, an assumption that a component works properly may have to be revised to false to explain a fault. Another subset of antecedents consists of *facts*, which are always true.

We refer to a set of assumptions as an *environment*. Each node in the ATMS is marked with a set of environments (any one justifies the node), referred to as *labels*. Each label is consistent and minimal. A label is consistent if all its environments are consistent. A label is minimal if no environment of the label is a superset of any other in that label. Inconsistent environments are called *conflicts*. A label of an assumption is a set of a set of the assumption itself. For example, label of the assumption  $B$  is  $\{\{B\}\}$ . A label of a fact consists of one empty environment. For example, the label of any fact is  $\{\{\}\}$ .

When a new justification is added, the ATMS computes a new label for its consequent node, and if there is a change in the new label, the new label is propagated through the network of justifications. The label of the node with respect to the new justification is the cross product of the labels of its antecedents. If a node has more than one justification, its label is the union of the labels contributed by those justifications. Consider the example in figure 5. The label of  $n1$  due to  $j1$  is  $\{\{A, B\}\}$ , and due to  $j2$  is  $\{\{D, E\}\}$ . The combined label due to  $j1$  and  $j2$  is  $\{\{A, B\}, \{D, E\}\}$ . The environment  $\{C, D, E\}$  has been removed from the label of  $n2$  because it is a superset of the conflict  $\{C, D\}$ . The final label of  $n2$  is  $\{\{A, B, C\}\}$ . The label of the contradiction node  $n3$  is  $\{\}$ .

#### 4.2 A model-based diagnosis example

Consider the circuit in figure 6, which consists of two exclusive OR gates X1, X2, two AND gates A1, A2 and an OR gate O1. Assume the inputs are  $A = 1$ ,  $B = 1$  and  $C = 0$ . On the correct working of the circuit, the other values are  $X = 0$ ,  $Y = 0$ ,  $Z = 1$ ,  $F = 0$ , and  $G = 1$ . Now the outputs are measured showing  $F = 1$  and  $G = 0$ . From the measurements, it is possible to deduce that at least one of the following sets of

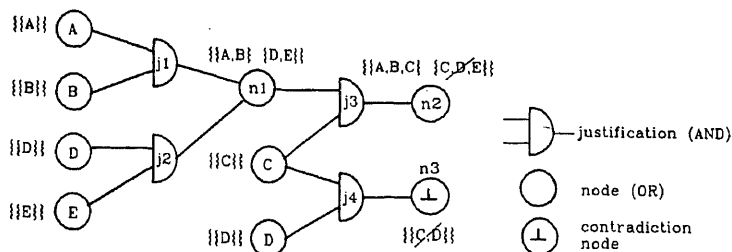
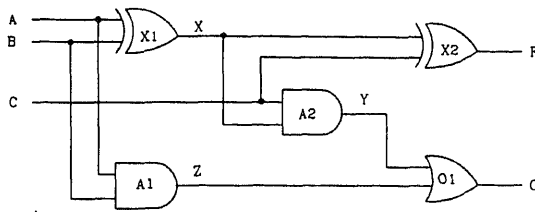


Figure 5. A dependency network showing nodes and justifications.



**Figure 6.** An example (full adder) for model-based diagnosis.

components is faulty:

$$[X1, A1], [X1, O1], [X2, O1], [X2, A1].$$

More measurements can isolate the faulty components. In the following discussions we assume that the interconnections are intact and hence they are not modelled explicitly. But they can be modelled as components with input equal to the output.

**4.2a Representing the system and its behaviour:** As mentioned earlier, the model of the system is a description of its physical structure plus models for each of its constituents. Each component is modelled as a set of constraints. A constraint is a relationship between variables. Consider the example in figure 6. A constraint  $A \oplus B \Rightarrow X$  can be defined to derive  $X$  from  $A$  and  $B$ . If the values for  $A$  and  $B$  are known, the constraint fires and produces a value for  $X$ . For example, if  $A = 1$  and  $B = 0$ , then the constraint produces  $X = 1$ . Similarly,  $F$  can be derived using the constraint  $X \oplus C \Rightarrow F$ . If we define both the constraints  $A \oplus B \Rightarrow X$  and  $X \oplus C \Rightarrow F$ , the interconnection between  $X1$  and  $X2$  is automatically satisfied. If the values for  $A, B$  and  $C$  are known, then the first constraint (for  $X1$ ) produces a value for  $X$  and the second constraint (for  $X2$ ) produces a value for  $F$ . The constraints described above for  $X1$  and  $X2$  form the partial models for the components  $X1$  and  $X2$  (the complete models are described in the next subsection). Once the complete models for all the components are defined, the description of the device must be complete.

**4.2b Deriving the behaviour:** The behaviour of the system is generated by a method called *constraint propagation*. Constraint propagation operates on variables, values and constraints. Given a set of initial values, constraint propagation assigns each variable a value that satisfies the constraints. If sufficient values are known for a constraint, that constraint triggers and produces a new value, which in turn may trigger other constraints and so on.

The constraint propagation process derives new values by propagating known initial values through a set of constraints (which represents a set of components). Each derivation is recorded in the ATMS with its dependencies, which trace out a particular path through the constraints that the inputs have taken. For example, if  $A = 1$  and  $B = 1$ , and  $A1$  is working correctly (figure 6), then the constraint propagation produces  $Z = 1$  and adds a justification:  $A = 1, B = 1, A1 \Rightarrow Z = 1$ , where  $A1$  is the assumption about the AND gate  $A1$ , which shows that  $Z = 1$  depends upon the correct functioning of  $A1$ . If we assume that  $A = 1$  and  $B = 1$  are facts, then  $Z = 1$  receives a label  $\{A1\}$ . One advantage with this approach is that it does not differentiate between inputs and outputs. A path may begin at any point in the circuit where a measurement has been taken. Also, it is not necessary to make any assumption about



the direction of the signal flow. As a result, the component model should capture all the constraints such that any terminal (variable) value must be derivable, provided some other values are known. The complete model for the example in figure 6 is shown below.

A1:

$$A \wedge B \Rightarrow Z, Z = 1 \Rightarrow A = 1 \wedge B = 1 \\ Z = 0 \wedge A = 1 \Rightarrow B = 0, Z = 0 \wedge B = 1 \Rightarrow A = 0$$

A2:

$$X \wedge C \Rightarrow Y, Y = 1 \Rightarrow X = 1 \wedge C = 1 \\ Y = 0 \wedge X = 1 \Rightarrow C = 0, Y = 0 \wedge C = 1 \Rightarrow X = 0$$

X1:

$$A \oplus B \Rightarrow X, B \oplus X \Rightarrow A, A \oplus X \Rightarrow B$$

X2:

$$X \oplus C \Rightarrow F, F \oplus X \Rightarrow C, F \oplus C \Rightarrow X$$

O1:

$$Y \vee Z \Rightarrow G, G = 0 \Rightarrow Y = 0 \wedge Z = 0 \\ G = 1 \wedge Z = 0 \Rightarrow Y = 1, G = 1 \wedge Y = 0 \Rightarrow Z = 1$$

Let us assume that the initial values are  $A = 1$ ,  $B = 1$  and  $C = 0$ . These values are stored as facts within the ATMS. Propagation produces  $X = 0$ ,  $Y = 0$ ,  $F = 0$ ,  $Z = 1$  and  $G = 1$ . Every derivation is stored in the ATMS, and is associated with a label  $\{e_1, \dots, e_n\}$ , where  $e_i$ 's are environments. The label shows from which set of components the value has been derived. In the following discussions we denote an assertion  $x$  with the supporting environments,  $e_1, e_2, \dots$  as  $[x, e_1, e_2, \dots]$ . Now the database contains the following:

$$\begin{array}{ll} [A = 1, \{ \}] & [B = 1, \{ \}] \\ [C = 0, \{ \}] & [X = 0, \{X1\}] \\ [Y = 0, \{X1, A2\}] & [Z = 1, \{A1\}] \\ [F = 0, \{X1, X2\}] & [G = 1, \{A1, O1\}] \end{array}$$

**4.2c Conflict detection:** As shown in the previous section each derivation is labelled with a set of environments, where an environment is a set of correctness assumptions of the components. For example, the label  $\{X1, X2\}$  of  $F = 0$  shows that  $F = 0$  is derivable only when the components X1 and X2 are working correctly. If the observed value differs from the predicted value, then the set of environments of the predicted value becomes inconsistent. An inconsistent environment is called a conflict. We denote a conflict as  $\langle A_1, A_2, \dots, A_n \rangle$  where  $A_i$ 's are assumptions. For example, if  $F$  is measured to be 1 then  $\{X1, X2\}$  becomes a conflict,  $\langle X1, X2 \rangle$ . It shows that X1 or X2, or both could be faulty. Since the label for a datum is minimal, we record only minimal conflicts, and its supersets are not explored.

**4.2d Candidate generation:** A candidate is a particular hypothesis which explains how the actual artifact differs from the model. A candidate is represented by a set of assumptions, indicated by  $[A_1, A_2, \dots]$ . Candidates have the property that any superset of a possible candidate for a set of symptoms must be a possible candidate as well. Thus the candidate space can be represented by a set of minimal candidates. The goal of candidate generation is to identify the complete set of minimal candidates.

Given no measurements, every component must be working correctly, and the single minimal candidate is [ ].

Candidates are generated from the conflicts. Whenever a new conflict is discovered, any previous minimal candidate which does not explain the new conflict is replaced by one or more superset candidates which are minimal, based on this new information. This is accomplished by replacing the old minimal candidate with a set of new tentative minimal candidates each of which contains the old candidate plus one assumption from the new conflict. Any tentative new candidate which is subsumed or duplicated by another is eliminated, the remaining candidates are added to the set of new minimal candidates.

Consider the example in figure 6. Initially there is no conflict, thus the minimal candidate is [ ]. The resulting database when  $A = 1$ ,  $B = 1$  and  $C = 0$  is as shown in §4.2b. If  $F$  is measured to be 1, then  $[F = 1, \{ \}]$  is added to the database ( $F = 1$  is a fact). Propagation produces  $[X = 1, \{X2\}]$  and  $[Y = 0, \{X2, A2\}]$ . The inconsistency between  $[F = 1, \{ \}]$  and  $[F = 0, \{X1, X2\}]$  produces a new minimal conflict  $\langle X1, X2 \rangle$ . The minimal candidates are  $[X1]$  and  $[X2]$ . Next suppose we measure  $G$  to be zero. Propagation gives  $[Y = 0, \{O1\}]$  and  $[Z = 0, \{O1\}]$ . The symptom  $G = 0$  but not 1 produces the conflict  $\langle A1, O1 \rangle$ . The new minimal candidates are:

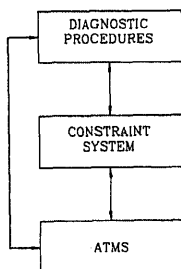
$[X1, A1], [X1, O1], [X2, A1], [X2, O1]$ .

The resulting database is shown below:

$[A = 1, \{ \}]$	$[B = 1, \{ \}]$
$[C = 0, \{ \}]$	$[F = 0, \{ \}]$
$[G = 0, \{ \}]$	$[X = 0, \{X1\}]$
$[X = 1, \{X2\}]$	$[Y = 0, \{X1, A2\}, \{X2, A2\}, \{O1\}]$
$[Z = 0, \{O1\}]$	$[Z = 1, \{A1\}]$

Further, if  $X$  is measured to be 0, it produces the new minimal conflict  $\langle X2 \rangle$ . The propagation gives  $[Y = 0, \{A2\}]$ . The minimal candidates are  $[X2, A1]$  and  $[X2, O1]$ . Finally, measuring  $Z = 0$  produces the conflict  $A1$ , and the final candidate is  $[X2, A1]$ . The final diagnosis is  $[X2, A1]$ . The malfunctioning components  $A1$  and  $X2$  explain the symptoms  $F = 1$  and  $G = 0$ .

**4.2e Implementation:** The system is implemented using an ATMS and a constraint system. The resulting architecture is shown in figure 7. From the above discussions it is clear that the system has to record the inferences (behaviour) and the set of



**Figure 7.** The architecture of the model-based diagnosis system.

assumptions under which they are true. An inference may depend upon other inferences. Thus the system is dealing with a network of inferences and their dependencies. The ATMS is a good tool for recording these inferences and dependencies in an efficient way. The current implementation of the ATMS (Mahabala & Kurup 1991a) provides a flexible interface with the application program. Now the bulk of the problem solving can be organized within the ATMS using consumers (de Kleer 1986b; Dressler & Farquhar 1990), where a consumer is a piece of code attached to a node which does some problem-solving work at the node.

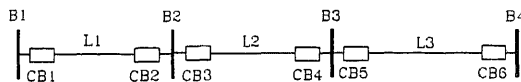
The constraint system (Kurup & Mahabala 1992) provides functions to represent definitions of components, to create instances of specific components, and to set values to variables. We can create a library of component definitions using the functions provided by the constraint system, and they can be loaded whenever the structure of the system needs to be defined. Whenever the structure changes, a component can be created appropriate for the structure from the library of definitions. Thus, this approach can save an enormous amount of work which would have been required if there was no facility for establishing a library of components. Consider the example in figure 6. Let us assume that a library of components for digital circuits has been already established. Now the structure of the circuit can be defined by creating the specific instances for the components X1, X2, A1, A2, and O1 from the library. When the constraint system runs, it records the structure and behaviour (in the form of consumers) in the ATMS. That is, finally everything boils down to the ATMS. The details of implementation of the constraint system can be found in Kurup & Mahabala (1992). The conflict detection algorithm is implemented as part of the constraint system. The algorithm runs whenever a new observation is added or a new value is derived for a variable.

The top level diagnostic procedures include the commands to create the description of the system to be diagnosed and the candidate generation algorithm. It attaches the candidate generation algorithm to an ATMS node (called FALSE, which denotes contradiction) as a consumer. Whenever a conflict (nogood) is detected, the consumer runs and takes the conflicting assumptions as argument and produces the candidates. More details on the encoding techniques are discussed in Mahabala & Kurup (1991b).

#### 4.3 *Diagnosis of a power system network. Practical application*

The purpose of a power system network is to distribute the electrical power from the generating sources to the customers in an efficient manner. The major components of the power system include the power sources (generators), transformers, connecting lines, buses, isolators and protective devices (circuit breakers, relays etc.). Normally the circuit breakers are closed and the lines carry power. When a fault occurs, the appropriate protective devices will operate and a portion of the network, which includes the fault, will be isolated. The effect of faults may disturb a large portion of the network or even the entire network. Problems can and do occur in protective equipment. These faults may cause tripping of many circuit breakers which results in a lot of information (messages) reaching a control centre. Automated fault localization based on the messages helps to restore the isolated network portions quickly and to ensure reliable distribution of power.

Since deriving rules from experience is very difficult, or even impossible, for a large network or a network which keeps on getting extended, it is necessary to use a model-based approach. A portion of a power system network (110/230 kV) is shown in



**Figure 8.** Portion of a 110/230 kV power system network.

figure 8. Presently we consider two types of protection: differential protection for buses, transformers and generators, and distance-relay protection for lines.

Assume that due to some disturbance (say a fault in B2) CB2 is tripped by the differential protection, and CB4 is tripped at distance level 2 (zone 2). The model-based approach produces the following diagnoses, one of which should be true.

- (1) [B2, CB3] – bus B2 and the breaker CB3 could be faulty,
- (2) [CB2, CB4] – breakers CB2 and CB4 could be faulty.
- (3) [CB2, L2, CB3] – breaker CB2, line L2 (25% from CB3), and CB3 could be faulty.

Using additional information from the relays, the diagnosis system would isolate one fault – [B2, CB3]. We have implemented a power system network fault diagnosis system and tested on an actual 110/230 kV network (Kurup & Mahabala 1993).

## 5. Qualitative model-based diagnosis

The model-based approach for diagnosis given in §4 can provide more information. However, the computing cost is high and extensive knowledge about the device states and history is also needed. In certain cases, the theory or the rigorous understanding of the mechanisms involved may be lacking. The qualitative model-based approach will be able to provide diagnosis with a less rigorous understanding of the system's state and behaviour. In a qualitative model, one represents the values and nature of time behaviour of variables in qualitative terms, such as low, high, slow, fast etc. The input-output behaviour is also expressed in a qualitative manner, such as, if input goes up, the output goes down etc.

A single qualitative model for troubleshooting in all situations may not work. For example, diagnosing a faulty hydraulic circuit with a molecular level qualitative model of flow of liquids is difficult. A molecular level qualitative model of flow of liquids might be the right level to analyse the flow of molecules through a pump, regulator etc., but does not provide a useful qualitative model for diagnosing the pump, regulator etc. Hence, there is need for multiple models of the same device at various levels with different simplifying assumptions and operating assumptions. For analysing a device represented by multiple models at different levels, we should have the means to select the right qualitative model in a given situation. Information derived at a more abstract level, if any, should be used efficiently for analysis at the detailed levels.

The characteristics/requirements of the diagnostic system are:

- the system must be able to accurately diagnose both the single and multiple fault cases;
- due to interactions in a device, a fault in one component can propagate to other components; the system should be able to use this information;
- the system must reason about time because much information can be deduced from the sequence and time of events;
- a number of components combine to form a feedback loop. If there is a fault in any of the components, the fault can be magnified or compensated due to interactions

in the loop. The system must precisely determine the component responsible for the failure.

The approach to qualitative model-based diagnosis followed is the abduction-based approach as opposed to the consistency-based approach usually followed in model-based diagnosis.

The system to be diagnosed,  $D$ , is given by the pair  $\langle \text{COMP}, \text{MODEL} \rangle$ .  $\text{MODEL}$  is a description of the structure and behaviour of the system  $D$ .  $\text{COMP}$  is the set of components of  $D$ . The set of parameters producing information about the specific case under investigation is given by the context,  $\text{CTXT}$ .

Given the observed behaviour of the system at various time points,  $\text{OBS} = \{\text{OBS}t1, \text{OBS}t2, \text{OBS}t3, \dots\}$ , the diagnostic problem involves determining the modes of the components explaining the observations in  $\text{CTXT}$ , i.e., an assignment  $W$  for  $\text{COMP}$  is to be found such that  $\forall \text{OBS}tx \in \text{OBS}, (\text{CTXT} \cup \text{MODEL} \cup W) \vdash \text{OBS}tx$ . The model,  $\text{MODEL}$ , is represented at multiple levels of abstraction and approximation.

A framework for diagnosis with multiple levels of abstraction and approximation is presented in the following sections. Section 5.1 describes the architecture of the diagnostic system. Section 5.2 discusses an example, § 5.3 presents the multi-level qualitative model of the system described in § 5.2 and the multi-level qualitative reasoning. Section 5.4 gives a test case in which the solution to the diagnosis problem is given, which in turn gives the solution to the control problems of efficiency and selection in multi-level qualitative reasoning. Section 5.5 discusses the advantages of multi-level qualitative reasoning and § 5.6 gives the implementation status.

### 5.1 Architecture of the qualitative model-based diagnostic system

The architecture of the qualitative model-based diagnostic system is shown in figure 9. The justifications of the observations are found by multi-level qualitative reasoner

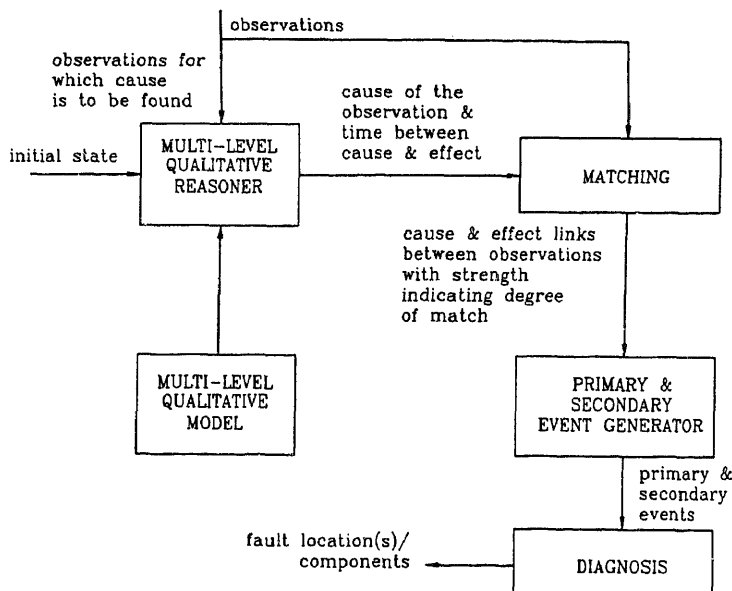


Figure 9. Architecture of the qualitative model-based diagnosis system.

using the behaviour generated from the multi-level qualitative model. The timing between the cause and the effect will also be generated. The justification for the observation derived from the qualitative model is verified among the set of observations for its presence by the matching module. The timing between the cause and effect derived from the qualitative model and the timing between the cause and effect actually observed are matched and a link strength based on the degree of match is formed between the cause and the effect by the matching module. The primary and secondary event generator generates the primary and secondary events based on the link strength. The diagnosis module generates the component(s)/location(s) responsible for the faulty behaviour from the primary and secondary events.

## 5.2 Example

The hydraulic system shown in figure 10 will be used to explain the concepts. The structural configuration of the system is as follows.

- pump with input *a* and output *b*;
- regulator with input *b1* and output *c*;
- pipe between *b* and *b1*;
- filter with input *c1* and output *d*;
- pipe between *c* and *c1*;
- electro hydraulic (EH) valve with inputs *d1* and *d2* and outputs *e* and *f*;
- pipe between *d* and *d1*;
- cylinder with input *e1* and output *e2*;
- pipe between *e1* and *e*;
- pipe between *e2* and *d2*;
- pipe between *f* and sump;
- pipe between sump and *a*.

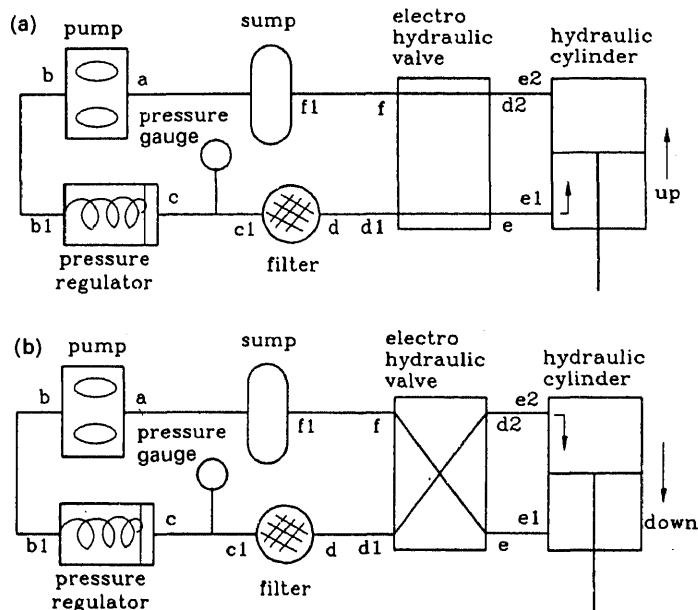


Figure 10. A hydraulic system during up (a) and down (b) of the EH valve.

Figures 10a and b show the system in two different positions (up and down traversal of piston rod) of the EH valve. The structural configuration for the system shown in figure 10b is the same as that for the system shown in figure 10a except that the EH valve has inputs  $d1$  and  $e$  and outputs  $d2$  and  $f$ .

The functioning of the circuit is as follows: Assume the EH valve is in the up-position as shown in figure 10a. The pump outputs the liquid at some pressure. The pressure regulator delivers liquid at a constant pressure irrespective of changes in pressure at its input (namely pump output). Any dirt particles present in the liquid are removed by the filter. The liquid then passes through the valve and reaches the cylinder. Due to the pressure differential across the two sides of the piston head in the cylinder, the piston head is lifted up and hence the load attached to the piston rod is lifted up. If the EH valve is in the down-position as shown in figure 10b, the piston head and the load attached to the piston rod are pushed down. Direction of piston movement (up or down) is controlled by the position of the EH valve. The fluid out of the cylinder empties into the sump.

Assume that there is a leak in the connecting pipe  $d - d1$ . There is a reduction in pressure that reaches the EH valve. This in turn causes a reduction in pressure that reaches the cylinder. This causes the load attached to the piston rod of the cylinder to go up very slowly or come down slowly depending on the position of the EH valve. The piston rod also does not move by the desired distance.

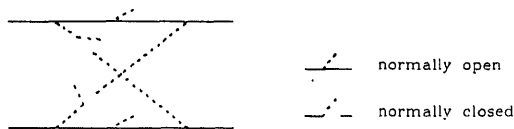
The reason for the reduction in the distance moved by the piston rod of the cylinder is due to a reduction in the pressure that reaches the cylinder. The reason for the reduction in pressure that reaches the cylinder is due to reduction in pressure that reaches the EH valve. There is no other justification for the pressure reduction at the input of the EH valve except for a leak in the connecting pipe  $d - d1$  or a fault in the filter or a fault in the filter and a leak in the connecting pipe  $d - d1$ . Incidentally, this is the way an engineer would reason about the working of the correct and faulty systems in qualitative rather than quantitative terms.

The change in value of pressure at  $e$  could have been due to a fault either in the filter or in the pipe or both in the filter and the pipe. A component behaves in a faulty manner when its mode is not as desired. The time taken for a pressure signal to traverse from the input to the output of the component in various modes for both the filter and the pipe are inferred from the model. The time taken for a pressure signal to travel from the filter to the EH valve with the filter in the normal mode and the pipe in the abnormal mode matches with the timing observed. The connecting pipe  $d - d1$  is identified as faulty.

### 5.3 Multi-level qualitative reasoning

In our system, the multiple models are based on the following abstractions:

- several components are combined into one. For example, even though a pump contains a lot of subcomponents, it can be considered as a component with a particular gain function in a certain level of analysis (or reasoning);
- the components are represented at different levels of detail. For example, a hydraulic circuit can be represented at three levels of detail. (i) Functional level where the valves are considered as switches which can be turned off or on, (ii) timing level where the components are modelled as a fluid resistance and fluid capacitance pair, (iii) flow level where the components and the interconnections are modelled in greater detail.



**Figure 11.** Figurative description of a qualitative functional model of the EH valve.

The hydraulic system shown in figure 10 will now be modelled at the three levels.

*Functional level:* At the functional level, the sub-components are combined and considered as a component with a particular gain function. The purpose of the model at this level is to check if the functioning of the functional blocks in the device is as expected. This checking is done considering the functional blocks in all possible modes. The general qualitative model of a component,  $C$ , at this level is:

status [relations], where status gives the mode of the component and the relations give the gain function.

The behaviour is analysed in terms of signals with order of magnitude values with the following quantity space  $[0, \text{LOW}, \text{MEDIUM}, \text{HIGH}, \text{INF}]$ .

The EH valve is modelled as a collection of simple valves (normally closed and normally open) that can be turned on or off. The qualitative functional model of the EH valve is shown in figure 11. The modes of the EH valve are as shown in figures 10a and b. This model helps to answer questions regarding the liquid reaching the cylinder.

The pump can be modelled as a component with a gain function. For a given input-pressure,  $\text{output-pressure} = (\text{input-pressure} + \text{pump-gain})$ . For example, if input-pressure is LOW and pump-gain in MEDIUM, then  $\text{output-pressure} = (\text{LOW} + \text{MEDIUM}) = \text{HIGH}$  (derived). There are three modes of the pump where the gain is correspondingly zero, between zero and maximum, and maximum. The arithmetic followed is a simple order-of-magnitude arithmetic (Raiman 1986).

The cylinder can be modelled as a component with a gain function, where the distance moved by the piston rod is  $(\text{input-pressure} \times \text{cylinder-gain})$ . The cylinder-gain depends on the distance available for the piston to move. The five modes of the cylinder are open up  $\uparrow$ , open down  $\downarrow$ , working up  $\uparrow$ , working down  $\downarrow$  and stuck. The distance available for the piston to move is maximum in the first two modes, between zero and maximum in the next two modes and zero in the stuck mode.

The regulator can be modelled as a component with a gain function, where the  $\text{output-pressure} = (\text{flow-rate} \times \text{regulator-gain})$ . The regulator-gain depends on the area available for the flow. The three modes of the regulator are open, closed and working. The area available for the flow in the above modes are maximum, zero, and between zero and maximum, respectively.

The filter can be modelled as a component, where the  $\text{output-pressure} = (\text{input-pressure} \times \text{filter-gain})$ , where the gain depends on the area available for the flow. The three modes of the filter are open, closed and working. The area available for the flow in the above modes are maximum, zero, and between zero and maximum, respectively.

*Timing level:* At the timing level, the changes are that the components are more detailed with each component replaced by an RC pair (fluid resistance and fluid capacitance) in various modes of the components. The behaviour analysis involves two parameters  $R$  and  $C$ .

The purpose of the model at this level is to compute the time delay and check if



the observed delay between two components is as expected. Delay depends on the time constant of the path established by functional-level analysis. The time constants are calculated using the fluid resistance and the fluid capacitance in the path.

*Flow level:* At the flow level, the components and the interconnections are modelled in greater detail and the behaviour is analysed in terms of pressure and flow-rate. The purpose of the model is to check whether the components in the functional blocks behave as expected. The general qualitative model of the component *C* at the flow level is

modality [conditions] relations,

where the modality of a component is open, closed, working etc., and the [conditions] express the conditions that are true in the particular mode and the relations are the equations that hold in that mode.

There will be some bridging relations between the components at the functional level and the components at the flow level. The model of the bridging relation is expressed as:

component at the functional level;  
components at the flow level;  
relations.

For example, the filter and  $d - d1$  pipe subsystem forms the functional level component, and its corresponding components at the flow level are the filter and the  $d - d1$  pipe. Some of the relations existing between them are:

- the input of the component consisting of the filter and  $d - d1$  pipe subsystem at the functional level is the input of the filter at the flow level;
- the output of the component consisting of the filter and  $d - d1$  pipe subsystem at the functional level is the output of the  $d - d1$  pipe at the flow level.

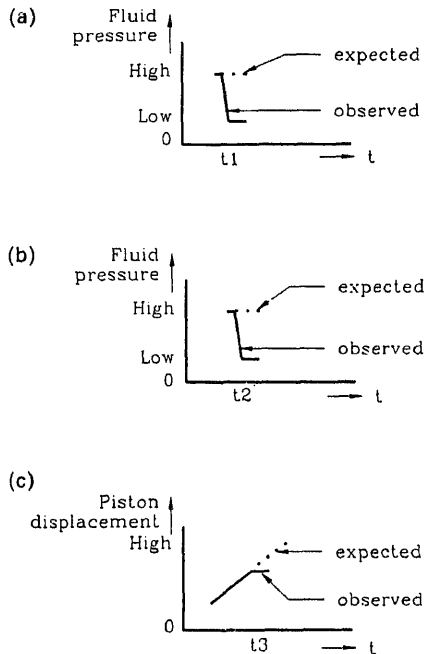
Given a domain described at multiple levels of abstraction, the approach taken is to start behaviour generation at the most abstract level and move down to a deeper level of reason only if the diagnostic problem demands it. Certain faults can also force the device into new modes of operation.

The algorithms for the reasoning at any level are given in Arun Kumar & Mahabala (1992). Several approaches to qualitative reasoning with multiple models have been made (Collins & Forbus 1987; Falkenhainer & Forbus 1988; Addanki *et al* 1989; Hibler & Biswas 1989; Liu & Farley 1990). Several approaches to qualitative reasoning with a single model have been that of de Kleer & Brown (1984), Forbus (1984) and Kuipers (1986).

#### 5.4 Test case

The following test case illustrates the solution to the diagnosis problem using reasoning with multiple models. Consider the changes observed in the system behaviour over a period of time, which is shown in figure 12. In figure 12,  $(t2 - t1)$  is the observed time delay in the signal flow from  $d1$  to  $e1$  and  $(t3 - t2)$  is the observed time delay between the change at  $e1$  and the change at the cylinder output.

The possible justifications for the observation at  $t3$  (figure 12c) as generated by the QUALITATIVE REASONER (figure 9) at the functional level can be one of the following:



**Figure 12.** Observations of deviations in pressure at  $d1$  (a) and  $e1$  (b) and piston displacement (c).

- (1) fall in pressure at  $e1$  in mode open  $\uparrow$  of the cylinder at an earlier instant of time;
- (2) fall in pressure at  $e2$  in mode open  $\downarrow$  of the cylinder at an earlier instant of time;
- (3) fall in pressure at  $e2$  in mode working  $\downarrow$  of the cylinder but slower than (2) at some earlier instant of time;
- (4) fall in pressure at  $e1$  in mode working  $\uparrow$  of the cylinder but slower than (1) at some earlier instant of time.
- (5) fall in flow-rate through the cylinder at some earlier instant of time.

The MATCHING module finds a change in pressure at  $e1$  at  $t2$  (figure 12b) among the set of observations. The cause for the change observed at  $t3$  (figure 12c) could be one of (1) or (4). To select between (1) and (4) for the cause of the change observed at  $t3$  (figure 12c), the timings in both the modes are generated at the timing level. The time taken for the pressure signal between cylinder input and output in mode (1) calculated using the model at the timing level and the observed timing between  $e1$  and  $e2$  match. Therefore, change in pressure at  $e1$  at  $t2$  (figure 12b) is established as the cause for the observation at the output of the cylinder at  $t3$  (figure 12c) by the PRIMARY AND SECONDARY EVENTS GENERATOR (figure 9).

The possible justifications for observation at time  $t2$  (figure 12b) generated by the QUALITATIVE REASONER at the functional level can be one of the following.

- (1) Fall in pressure at  $d1$  in the UP mode shown in figure 10a at some earlier instant of time;
- (2) Fall in flow-rate across the EH valve at some earlier instant of time in the closed mode.

The MATCHING module finds the observation at  $d1$  at time  $t1$  (figure 12a). The timing between  $d1$  and  $e1$  calculated by the model at the timing level and the observed

timing between  $d1$  and  $e1$  also match. Therefore change in pressure at  $d1$  at  $t1$  (figure 12a) is established as the cause for the observation at the output of EH valve at  $e1$  at  $t2$  (figure 12b) by the PRIMARY AND SECONDARY EVENTS GENERATOR (figure 9).

The possible justification for observation at time  $t1$  (figure 12a) generated by the QUALITATIVE REASONER at the functional level can be one of the following.

1. Fall in flow-rate through the filter at some earlier instant of time;
2. fall in pressure at  $c1$  in mode open of the filter at some earlier instant of time;
3. fall in pressure at  $c1$  in mode working  $\uparrow$  of the cylinder but slower than 2 above at some earlier instant of time.

The MATCHING module does not find any of the above in the set of observations. Hence, there is no justification for the observation at time  $t1$  (figure 12a). The fault is isolated to the subsystem consisting of the filter and connecting pipe  $d - d1$ .

Now, the fault has to be further localised. With the nominal values at  $c1$  and  $d1$ , the modes of the components and the possible values at the intermediate points are generated by the QUALITATIVE REASONER at the flow-level. With the changed value at  $d1$  and the nominal value at  $c1$ , the possible modes of the components and the possible values at the intermediate points are generated by the QUALITATIVE REASONER at the flow-level.

The modes which explain the change in the value at  $d1$  are:

- (1) the pipe is leaky;
- (2) the filter is clogged; or
- (3) the pipe is leaky and the filter is clogged.

Given a small change in pressure at the filter input, an observation is made on the time taken for the pressure change to reach the valve input. This is checked with the timing calculated from the model at the timing level in each of the three modes given above. The timings calculated from the model at the timing level in the case of the pipe, being leaky, match the observed timings. Thus the fault has been localised to the pipe being leaky. The diagram of the multi-level simulation process is shown in figure 13.

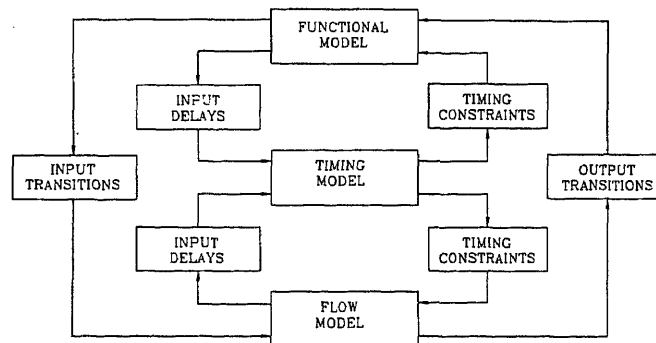


Figure 13. Interactions between models at different levels.

### 5.5 Discussion

It is difficult to simulate the circuit and test for all the three failure modes simultaneously. At the functional level, the failures associated with the functioning of each of the functional blocks are identified. After isolating the fault to a functional block, the fault location is localised to the subcircuits at the flow level in that block. Any ambiguity as to which of the components in the functional block is responsible while localising the fault at the flow level is resolved at the timing level. Thus, one of the advantages of using qualitative models at multiple levels for generation of the system behaviour is that the information derived at the higher levels is used for behaviour generation at more detailed levels, thereby making the analysis computationally less expensive.

In our framework, new knowledge is added by designing an appropriate new model which will work with the existing ones. For example, a timing model can be added to the functional model to increase the completeness of the description.

### 5.6 Implementation

Prototype Lisp programs have been written for simulation at functional, timing and flow level.

## 6. Conclusion

Diagnosis is a value-added area for the application of expert system technology. The traditional symptom-and-cause approach can be captured by the diagnosis tree or a set of rules. Required knowledge is based on experience and is referred to as causal model. One can expect wide use of the causal approach in years to come. The inflexibility of the causal model in terms of not being able to take care of even minor variations in the configuration can be overcome by the model-based approaches. The qualitative approach for describing behaviour, which often is the basis of design, can also be used to build a diagnostic system. Mechanical systems are generally handled by the use of qualitative models. The need for greater flexibility is provided by the model-based approach. One can handle a wide range of systems built using a set of components whose models are stored in a library. Power systems are suited for the model-based approach. Just as an engineer would first try to locate faults based on symptoms (causal model) and move to qualitative reasoning for more complex faults and finally resort to analytical approaches (model-based), so also one can build a multi-model diagnosis system which localises faults using causal models and see if faults can be further localised by qualitative models, and lastly resort to analytical models if warranted.

The authors would like to thank the Department of Electronics, Government of India and the United Nations Development Programme (UNDP) for assistance in carrying out this work.

**Appendix A. A sample diagnosis session for a voltage stabilizer (see § 2.3):**

Please input your name: A Raman  
Please input your code: 12n45  
Your competence is: 90  
Please enter the Territory code: 300  
Please enter the Customer Name: KBCS LAB  
Please enter the Customer address: IIT MADRAS 600 036  
Please enter the Machine serial Number: 320  
Please enter the Machine Rating (KVA/Phase): 430  
Please enter the Model Name: SCVS, KRYKARD

(IITMDESS) starts

Is there any problem in the system?  
(yes/no/suspend): yes

Is there output from stabilizer and any panel indication?  
(yes/no/suspend): yes

Does the panel show permanent indication of output-low or output-high but not input-high or input-low?  
(yes/no/suspend): no

Does the panel show permanent indication of input-low or input-high, and output-low or output-high?  
(yes/no/suspend): yes

Likely cause: Variac brush broken or contacts defective.  
{the system will find that the likely causes are not "broken brush" or "faulty pressure plate" of variac. Continuing with the diagnosis...}

Likely cause: Cut in the variac turns.  
Check: With DMM set to 1000 V AC range, check voltage between Line and Neutral, Line and Fixed Tap, Neutral and Fixed Tap.  
Type 'done' when operation is complete,  
'help' for help or 'suspend' to suspend diagnosis: done  
Input integer value for Vnf: 230  
Input integer value for Vlf: 120  
Input integer value for Vin: 230

REPAIR to be done is the following

There is a cut in turns between Neutral and Fixed Tap  
Replace Variac.  
Type 'done' when repair is over  
'help' for the help message or 'suspend' to suspend diagnosis: done

**OLD PART REPLACED:**

Name of the part: Variac  
Type: 2  
Serial No: 120

**NEW PART REPLACED:**

Name of the part: Variac  
Type: 2  
Serial No. 127

Does the panel show permanent indication of input-low or input-high, and output-low or output-high?  
(yes/no/suspend): No

Is there any problem in the system?  
(yes/no/suspend): No

Diagnosis Over  
Please input your remarks for the current session:  
VARIAC IS REPLACED.

**Appendix B. A portion of the rule-base for the PC diagnosis (§ 3):**

Rule for checking floppy is:

If there is a floppy disk problem  
And the disk reads correctly  
And the disk does not write  
Then check other aspects

The rules for checking other aspects are given below.

R1:

If disk does not appear ok  
Then the problem is due to physical damage of the disk.

R2:

If disk does appear ok  
And the write protect hole on the disk is covered  
Then problem is due to the write protect switch.

R3:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem does not persist after removing the resident programs  
Then problem is due to the resident programs.

R4:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem persists after removing the resident programs  
And the system has a second disk drive or access to another compatible computer

And the disk works the same way on the other drive  
Then problem is due to a bad disk.

R5:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem persists after removing the resident programs  
And the system has a second disk drive or access to another compatible computer  
And the disk does not work the same way on the other drive  
And the disk loads properly and the motor turns after cleaning the read head  
Then problem is due to dirt on the read head.

R6:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem persists after removing the resident programs  
And the system has a second disk drive or access to another compatible computer  
And the disk does not work the same way on the other drive  
And the disk does not load properly and the motor does not turn after cleaning the read head  
Then problem is due to bad floppy drive.

R7:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem persists after removing the resident programs  
And the system does not have a second disk drive or access to another compatible computer  
And the disk loads properly and the motor turns after cleaning the read head  
Then problem is due to dirt.

R8:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem persists after removing the resident programs  
And the system does not have a second disk drive or access to another compatible computer  
And the disk does not load properly and the motor turns after cleaning the read head  
And it happens with only one program  
Then problem is due to difficulty of some programs working with some disks.

R9:

If disk does appear ok  
And the write protect hole on the disk is not covered  
And problem persists after removing the resident programs  
And the system does not have a second disk drive or access to another compatible computer  
And the disk does not load properly and the motor turns after cleaning the read head  
And it does not happen with only one program  
And it happens with one particular floppy  
Then problem is due to a bad disk or a particular brand of disks.

R10:

If disk does appear ok

And the write protect hole on the disk is not covered

And problem persists after removing the resident programs

And the system does not have a second disk drive or access to another compatible computer

And the disk does not load properly and the motor turns after cleaning the read head

And it does not happen with only one program

And it does not happen with one particular floppy

Then problem is due to bad floppy drive.

*Typical diagnosis session:* Let us assume that the floppy disk problem is due to a bad disk. Now a possible diagnosis session would be the following:

Q: Is there any floppy disk problem? (y/n)

A: y

Q: Does the disk read correctly? (y/n)

A: y

Q: Does the disk write correctly? (y/n)

A: n

(Now the rules for other aspects will be checked).

Q: Does the disk appear ok? (y/n)

A: y

(Now R1 fails and R2 is selected).

Q: Is the write protect on the disk covered? (y/n)

A: n

(Now R2 fails and R3 is selected).

Q: Remove all the resident programs if any.

Does the problem persist? (y/n)

A: y

(Now R3 fails and R4 is selected).

Q: Do you have a second disk or access to another compatible computer? (y/n)

A: y

Q: Does the disk work in the same way as in the other drive? (y/n)

A: y

Now R4 is confirmed and the diagnosis is "the bad floppy disk."



## References

- Addanki S, Cremonini R, Penberthy J S 1989 Reasoning about assumptions in graph of models. *Proc. Int. Joint Conf. Artif. Intell.* (Detroit, MI) 2: 1432-1438
- Arun Kumar A T, Mahabala H N 1992 Multi-level qualitative reasoning applied to hydraulic circuits. Technical Report, Dept. of Comput. Sci. & Eng., Indian Inst. Technol., Madras
- Collins J, Forbus K D 1987 Reasoning about fluids via molecular collections. *Proc. Am. Assoc. Artif. Intell.* 87 (Seattle, WA) 1: 590-594
- de Kleer J 1986a An assumption-based TMS. *Artif. Intell.* 28(2): 127-162
- de Kleer J 1986b Problem solving with ATMS. *Artif. Intell.* 28(2): 197-224
- de Kleer J, Brown J S 1984 A qualitative physics based on confluences. *Artif. Intell.* 24(1-3): 7-83
- de Kleer J, Williams B C 1987 Diagnosing multiple faults. *Artif. Intell.* 32(1): 97-130
- Dressler O, Farquhar A 1990 Putting the problem solver back in the driver's seat: Contextual control of the ATMS. *Lecture notes in artificial intelligence* (eds) J P Martin, M Reinfrank (Springer-Verlag) 515: 1-16
- Falkenhainer B, Forbus K D 1988 Setting up large scale qualitative models. *Proc. Am. Assoc. Artif. Intell.* 88 (St. Paul, MN) 1: 301-306
- Forbus K D 1984 Qualitative process theory. *Artif. Intell.* 24(1-3): 85-168
- Fukul C, Kawakami J 1986 An expert system for fault section estimation using information from protective relays and circuit breakers. *IEEE Trans. Power Delivery.* PRWD-1(4): 83-89
- Hibler D L, Biswas G 1989 TEPS: The thought experiment approach to qualitative physics. *Proc. Int. Joint Conf. Artif. Intell.* 89 (Detroit, MI) 2: 1279-1284
- Kuipers B 1986 Qualitative simulation. *Artif. Intell.* 29(3): 289-338
- Kurup R R, Mahabala H N 1992 Implementation of a constraint system. *Proc. Int. Assoc. Sci. Technol. Dev., Int. Symp. Modeling, Identification and Control*, Innsbruck Austria
- Kurup R R, Mahabala H N 1993 Model-based diagnosis of power system network faults. Technical Report, Knowledge-Based Computer Systems Laboratory, Indian Inst. Technol., Madras
- Liu Z Y, Farley A M 1990 Shifting ontological perspectives in reasoning about physical systems. *Proc. Am. Assoc. Artif. Intell.* 90 (Boston, MA) 1:395-400
- Mahabala H N, Kurup R R 1991a Implementation of an extended ATMS. Technical Report, Knowledge-Based Computer Systems Laboratory, Indian Inst. Technol., Madras
- Mahabala H N, Kurup R R 1991b An implementation of a general diagnostic engine. *Proc. Int. AMSE Conference on Signals, Data & Systems* (Delhi: AMSE Press) 2: 137-150
- Mahabala H N, Ravikanth G 1990 IITMRULE user's manual, KBCS Laboratory, Indian Inst. Technol., Madras
- Mahabala H N, Ravi Prakash G, Managoli V V 1992 IITMDESS: A fault-tree based diagnosis shell, Technical Report, KBCS Laboratory, Indian Inst. Technol., Madras
- Raiman O 1986 Order of magnitude reasoning. *Proc. Am. Assoc. Artif. Intell.* 86 (Philadelphia, PA) 1: 100-104
- Struss P 1988 A framework for model-based diagnosis. Siemens Report TR INF2 ARM-10-88, Munich
- Talukdar S N, Cardozo E, Perry T 1986 The operator's assistant - an intelligent expandable program for power system trouble analysis. *IEEE Trans. Power Syst.* PRWS-1(3): 182-187
- Vesonder G T, Salvatore J S, Zieliski J E, Miller F D, Copp D H 1983 ACE: An expert system for telephone cable maintenance. *Proc. Int. Joint Conf. Artif. Intell.* 83 Karlsruhe, Germany 2: 116-121
- Williams T L, Orgren P J, Smith C L 1983 Diagnosis of multiple faults in a nation-wide communications network. *Proc. Int. Joint Conf. Artif. Intell.* 83, Karlsruhe, Germany 2: 179-181

## A fuzzy expert system approach using multiple experts for dynamic follow-up of endemic diseases

APURBA BANERJEE, ARUN KUMAR MAJUMDER<sup>1</sup> and  
ANUPAM BASU<sup>2</sup>

Department of Computer Science & Engineering, Indian Institute of  
Technology, Kharagpur 721 302, India

E-mail: <sup>1</sup>anupam@cse.iitkgp.ernet.in; <sup>2</sup>akmj@cse.iitkgp.ernet.in

**Abstract.** In this paper, an architecture of LEPDIAG – a knowledge-based system for on-line diagnosis and for monitoring prognosis of leprosy is presented. The important features of LEPDIAG that have been detailed are a multiple expert environment, a homeostatic expert containing the model of immune reaction, a performance evaluator that can compare the observed signs and symptoms with those predicted by the homeostatic expert and a prognostic expert which optimizes the management schedule for the patients. The entire systems is built around a fuzzy expert-system building tool FEXT to deal with the imprecise knowledge.

**Keywords.** Homeostasis; idiotypic network; immunological reaction.

### 1. Introduction

The advent of expert systems has concretised the advantage of artificial intelligence, in the form of practical systems. Expert systems have been applied to a number of generic task domains such as diagnosis (Chilausky *et al* 1976; Davis *et al* 1977; Chandrasekharan *et al* 1979), planning (Martin 1977), design (McDermott & Steele 1981), mineral exploration (Hart *et al* 1978) etc. A good review of expert system organisation and its applicability in different domains is available in Stefic *et al* (1982). The boom of expert system activities in medical diagnosis was set off by the seminal work on MYCIN (Buchanan & Shortliffe 1984) and was followed by a number of important developments like INTERNIST, ONCOCIN, TEIRESIAS etc. (Buchanan & Shortliffe 1984). Medical diagnosis activities also led to interesting architectures (Clancey 1983; Davis 1983; Reggia & Tuhim 1985; Patil 1987; Sticklen 1987).

Our investigations in this area convinced us that the domain of medical diagnosis demands the use of multiple expert modules instead of a monolithic structure of the first generation expert systems. The task of medical diagnosis treads into a number of decision activities of varying nature, ranging from abductive diagnosis, tentative

prognosis as well as a deep analyses of the particular status of the homeostatic equilibrium. Hence, for proper diagnosis and also for a reasonable attempt to model the cognitive process of a medical expert, it is essential to have an environment, where each of these areas of expertise is captured and allowed to work in close coordination.

Another aspect closely associated with medical diagnosis in the uncertainty of knowledge and data. Different approaches for dealing with uncertainty in the medical domain has been proposed such as CF in MYCIN (Buchanan & Shortliffe 1984), belief function (Shafer 1976) etc. The imprecise correlation between signs and symptoms with disease can be elegantly captured by fuzzy reasoning technique (Zadeh 1983a). A number of medical expert systems have adopted fuzzy techniques (Albin 1975; Perez-Ojeda 1976; Adlassnig & Kolarz 1982, pp. 219-47; Cios *et al* 1991).

In this paper, we present the architecture and reasoning techniques of LEPDIAG – an expert system for dynamic follow-up of endemic diseases.

The problem that is often encountered in chronic endemic diseases, such as leprosy, tuberculosis, AIDS etc., is the changes in the symptom and sign complex of a disease over time. Due to inherent immunological reactions in the body, it is expected that the symptom pattern will undergo considerable changes with change in the state of a disease. The effect of such changes become more pronounced as a consequence of medication given in the early stages of the disease. The development of a diagnostic system which addresses this problem requires an adequate model of the dynamics of the symptom-disease evolution pattern. In this paper, we have represented the immunological information and its effect on symptom-disease relationship in the form of rules. This approach enables us to deal with the changes in symptom-disease relationship over time and to interact with the conventional rule-based approach to medical diagnosis. In our current discourse, we intend to deal with a typical endemic disorder – leprosy.

LEPDIAG provides a diagnostic environment, involving multiple expert modules working in close coordination, through a shared framebase. The knowledge base consists of fuzzy rules clustered in functional partitions. The inference machine supports fuzzy inferencing and is implemented through a fuzzy expert system building tool FEXT.

For a proper understanding of the problem domain, a brief introduction to the disease is presented in the next section.

## 2. Leprosy as a disease

The most remarkable feature of leprosy is the wide variation of its manifestation depending on the immunological status of the individual. In some patients it may be manifested with a single nerve involvement or a single skin patch, while in others it may produce diffuse involvement of skin with multiple nodules together with polyneuritis and damage to the vital organs such as eyes, larynx, testes and bones. Every conceivable variation occurs between these two extremes. The causes of these variations lie in the variation of the patient's immune status and not merely on the bacterial strains with varying pathogenicity. In fact, it has been confirmed by Rees (1969) that leprosy bacilli from patients with different types of leprosy all behave in the same manner when injected into susceptible mice.

There are five clinical variations, seven histopathological and twelve immunohistological variations of the disease-complex (Ridley 1972). Each variation can be

denoted as a distinct disease-state within the whole range of the leprosy disease-spectrum. A single patient can pass through all the five clinical variations before getting cured, or while suffering from mutilating miseries on a finite time scale. Moreover, time-framed changes of clinical, pathological and immunohistological parameter values provide enough information to detect resistant infection, inadequate drug therapy, or immunological instability leading to various types of lepra reactions.

## 2.1 *Evolution of signs and symptoms over time*

Essentially, a set of signs and symptoms of a specific disease is true only at that time instance and depends only on the current status of the homeostatic stability of the patient's body (Ganong 1989, pp. 34–5). Homeostatic stability depends upon the quantum and quality of the environmental injury, i.e. on antigens entering into the system, type and quantity of drugs administered, inappropriate metabolism of proximate principles, stress and strains modulating genetic information flow. It also depends on the concerted attempt of the multiple subsystems, namely immune system, vascular system, endocrine system etc., to contain the behaviour within the viable range of dynamic equilibrium.

With reference to leprosy, duration of the treatment varies from a minimum of five years to the entire lifetime of the patient. According to the immunological status of the individual, a leprosy patient may swing in the spectrum of different disease-states corresponding to polar tubercular leprosy at one end to polar lepromatous leprosy at the other extreme (Nath 1983). The states of a disease and the changes in symptom pattern over time are primarily dependent on the immune reactions in the body. Hence, to capture the dynamic relationships between symptoms and diseases, we need to have a suitable model of the immune system. In the following subsection an attempt has been made to develop such a pragmatic model which is useful for our purpose.

## 2.2 *Immune reaction model*

Although several models of immunodynamics have been presented over the years (Richter 1978, pp. 219–27; Herzenberg & Black 1980), none of them is a useful tool to help in the diagnosis and prognosis of infective disorders. An interconnected idiotypic network model (Jerne 1973) can be accepted as a more feasible representative of the above class where the network provides control over the idiotypic interactions among the *T*-suppressors, *T*-helpers and the antibodies to any antigenic challenge, in any given immune system (Rich 1988). Details about the models can be found elsewhere (Richter 1978, pp. 219–27; Herzenberg & Black 1980; Perelson 1989). In the present work, the behaviour of the idiotypic network is represented in terms of rules. The rule-based approach has been adopted so that the immune reaction model can be easily integrated with the diagnostic expert system. Moreover, a rule-based approach, especially with fuzzy rules, can readily accommodate the imprecise measurement and estimation of antibody concentration and *T*-lymphocyte population.

For the proposed work, two hundred and fifty patients of leprosy in the School of Tropical Medicine, Calcutta have been examined. It is seen from the available data that though the *in-vivo* lepromin test results appear to be a weak parameter in determining a patients's clinical status, other *in-vitro* tests such as the LTT (lymphocyte

transformation test) etc. are potent enough in that respect. These observations indicate that the integral relationship among the antigenic load (BI), T-lymphocyte populations, index of T-lymphocyte suppression (LTT), and the level of antibodies is very irregular and complex. Available data also indicate that the manifestation of immunological parameters are time-dependent. Therefore, to represent the changes of the above parameters occurring over time and their corroboration with the manifested signs and symptoms, a rule-based model is necessary. An example rule of such a model is given below:

R1.

If

Antigen entering into the system is  $A$

AND T-S population is very low

AND T-H population is very high

Then

The quantity of the antibody  $Ab_x$  would be very low after a few days.

Such rules form the knowledge-base of the homeostatic expert HEXPERT, which has been entwined with a diagnostic expert system, a prognostic expert and a performance evaluator to form the composite system LEPDIAG. Thus it is possible to view the proposed system as a cooperating system of different knowledge sources, each dedicated to its own functional subtask. In the following section, we dwell on the detailed architecture of LEPDIAG and also describe the functions of each of the component expert systems. The interactions among these individual expert modules to form the composite whole of LEPDIAG is also explained.

### 3. Overview of the proposed system (LEPDIAG)

A schematic of LEPDIAG is shown in figure 1. The system consists of three expert system modules: diagnostic expert (DIAG), homeostatic expert (HEXPERT), and prognostic expert (PROG) together with a performance evaluator (PERF).

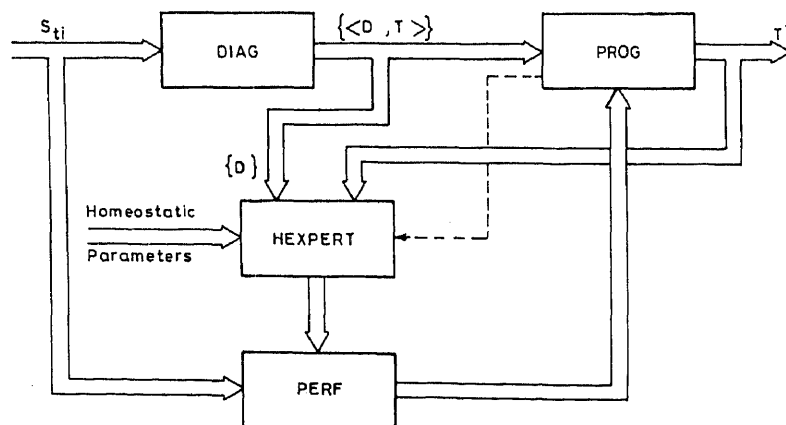


Figure 1. An overview of LEPDIAG.

**DIAG:** It is a diagnostic and therapeutic expert, which accepts the signs and symptoms of the patient, at any point of time, as the input. Suppose that signs and symptoms identified at time  $t$  are denoted by  $S_t$ . Based on the signs and symptoms  $S_{t_0}$  at time  $t_0$ , DIAG arrives at a list of possible diagnoses on utilizing the knowledge-base contained in it. Each identified disease is associated with a degree of belief and a list of candidate treatments. Thus the output of DIAG can be denoted by  $\langle D, T \rangle_{t_0}$  where  $D$  is a possible disease along with the associated belief and  $T$  is a list of treatments for this disease. Each treatment is also associated with a confidence factor.

**PROG:** The prognostic expert accepts as input the set of diseases and treatments  $\langle D, T \rangle_t$ , identified by DIAG at time  $t$  and utilises its prognostic knowledge-base to decide on the most likely disease and selects the most suitable treatment for it. The output  $T_{t_0}^*$  of PROG is fed to HEXPERT for further refinement of the prognosis as discussed below.

**HEXPERT:** The knowledge-base of the homoeostatic expert contains the knowledge about the "dynamic behaviour" (§ 2.1) of the respective disease. Given a set of possible diseases, administered treatments and the values of different homoeostatic parameters (such as lymphocyte count, bacterial index etc.) of the disease model, HEXPERT can predict some new signs and symptoms, which will be manifested after a predefined interval of time.

**PERF:** The performance evaluator module accepts inputs from two sources. The predicted set of signs and symptoms  $\hat{S}_t(D, T^*, P_h, t_0)$ , arrived at by HEXPERT is propagated to PERF. Also, as the patient visits the clinic at the next time instant (that is at time  $t_1$ ), the actual set of signs and symptoms  $S_{t_1}$ , is observed. This set ( $S_{t_1}$ ) forms the second input of PERF. The performance evaluator compares  $\hat{S}_t(D, T^*, P_h, t_0)$  and  $S_{t_1}$  and finds the extent of match.

In the next section the organisation of the knowledge bases of the different modules are discussed.

#### 4. Knowledge modules

The physical architecture of LEPDIAG is shown in figure 2. Different logical experts as shown in figure 1 are essentially a collection of knowledge modules partitioned in blocks of rule bases. Each block of knowledge module is dedicated to a particular phase of the entire task.

##### 4.1 DIAG

The first expert, the diagnostic and therapeutic expert (DIAG) is a constellation of seven knowledge modules. In the first phase, the module *KB1* is utilised to identify a set of possible diseases from a set of patient data, termed as complaint. Each of the diagnoses arrived at is associated with a degree of belief with a fuzzy truth value to indicate the relative possibility of different diseases. It is often necessary to carry out checks to find out some other components of complaints which might have been omitted by the patient. To this end, DIAG utilises *KB2* to suggest further checks

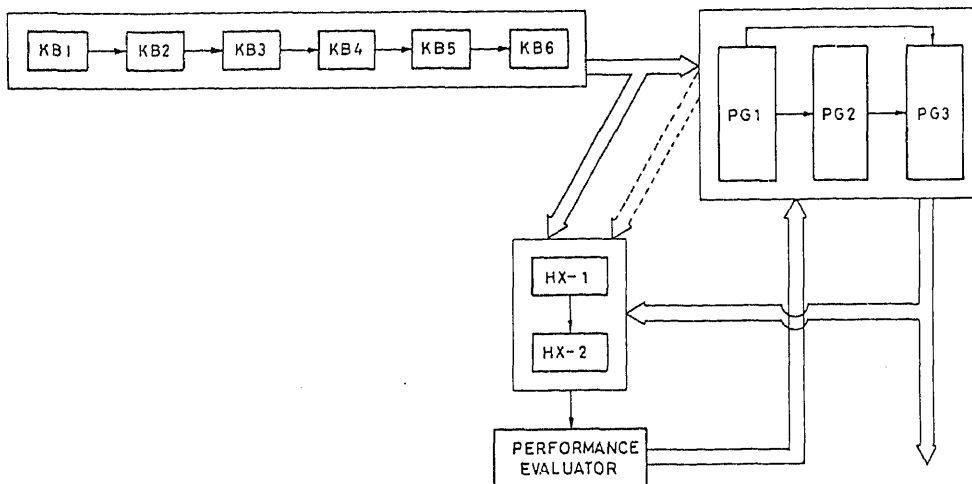


Figure 2. Physical architecture of LEPDIAG.

needed for the particular session. The check results thus obtained by the patient are utilised by the knowledge module *KB3* which generates a new set of facts with the modified belief of provisional diagnoses. The module *KB4* contains rules for suggesting special investigations like pathological and biochemical tests. There is a submodule *KB4'* within *KB4* which examines from the list of beliefs about the diagnoses whether any such special investigations would be necessary for the session or not. The test results obtained through *KB4* are utilised by the knowledge module *KB5*. This module basically analyses the investigation reports and modifies the associated belief of the provisional diagnoses accordingly, thus helping *DIAG* to arrive at a more confirmed set of provisional diagnoses. The task of diagnosis is basically abductive in nature. In LEPDIAG the reasoning is essentially forward chaining. However, the abductive nature of reasoning has been easily captured by forward chaining rules associated with belief measures. In abduction  $p \rightarrow q$  yields that if  $q$  is true then  $p$  is possibly true. In LEPDIAG such situations have been captured as  $q(\text{symptom}) \rightarrow p(\text{disease rate})$  (fuzzy belief). For the current session, *KB5* also checks whether the belief about leprosy is significant enough and invokes *KB6* if leprosy has been diagnosed, otherwise the expert system halts, printing the diagnosis having maximum belief. Knowledge module *KB6* enquires about some more clinical facts in terms of check-signs to classify the patient to the respective slot of disease-state from which the patient is suffering at that time of diagnosis. Since leprosy involves a spectrum of diseases, *KB6* determines belief about the variations of leprosy at the current-time instant. Thus *KB6* output consists of a set of leprosy-related diseases and their respective beliefs.

## 4.2 HEXPERT

The homoeostatic expert consists of two knowledge modules. The first module *HX1* gets the set of admissible diseases,  $D$  from *DIAG* and a single treatment regime  $T^*$  from the third expert *PROG* (discussed in §4.4). The knowledge-base *HX1* is utilised to identify the investigations to be done for the internal homoeostatic parameters ( $P_h$ ) (viz., immunohistological parameters in any infective disorder such as leprosy)

for each element of the set  $D$ . The second module  $HX2$  uses the values of the homoeostatic parameters ( $P_h$ ) along with the treatment  $T^*$  for simulating the homoeostatic process. This module also enquires about the time interval for which the simulation should be done for the respective  $D$ s. The module  $HX2$  calls the respective disease models for generation of predicted signs and symptoms with appropriate beliefs i.e.,  $\hat{S}_i(D, T^*, P_h, t_0)$  after the specified time interval.

#### 4.3 PERF – The evaluator

The set of predicted signs and symptoms released by HEXPERT is taken up by the evaluator (PERF) – a submodule for checking (comparison of) the match between the observed signs at time instant  $t(S_i)$  and the predicted ones i.e.,  $\hat{S}_i(D, T^*, P_h, t_0)$ . Comparison may lead to following two cases.

*Case I:* Suppose that the predicted and the observed set of signs and symptoms i.e.,  $\hat{S}_i(D, T^*, P_h, t_0)$  and  $S_i$  match. While computing the degree of match, fuzzy beliefs about the predicted and observed symptoms are taken into account. The evaluator compares a predicted symptom  $\hat{s}$  which belongs to  $\hat{S}_i(D, T^*, P_h, t_0)$  with a similar element  $s$  belonging to  $S_i$ . If for each  $\hat{s} \in \hat{S}_i(D, T^*, P_h, t_0)$ , there exists an  $s \in S_i$  such that the similarity measure (cf. § 5.3) about their belief is above the threshold, the PERF signals the “match” to PROG. But if the value of the similarity measure is less than the predefined threshold value then evaluator conveys the “mismatch” to PROG to initiate appropriate steps.

For example, in a case of borderline leprosy as a possible diagnosis, suppose that the initial signs and symptoms at time  $t_0$  (at time of first examination) are as follows:

- (1) number of skin patches: considerable;
- (2) shine over the skin lesions: moderate;
- (3) bilateral symmetry of the lesions: less marked;
- (4) area of loss of sensation: large.

Patients's information about immunohistological parameters are as follows:

- (1) bacterial index less than 3;
- (2) lepromin test 1+;
- (3) lymphocyte transformation test more than 3%;
- (4) antibody titre moderately high.

After treatment with Dapsone and Rifampicin for one month, the expected set of signs and symptoms ( $\hat{S}_{onemonth}$ ) are:

- (1) number of skin patches: small number;
- (2) shine over the skin lesions: absent;
- (3) bilateral symmetry: absent;
- (4) area of loss of sensation: moderately large.

When the patient returns after one month, the observed set of signs and symptoms are:

- (1) number of patches: moderately high;
- (2) shine over the lesions: less moderate;
- (3) bilateral symmetry: absent;
- (4) area of loss of sensation: large.



**Table 1.** Initial signs and symptoms and their fuzzy attributes.

Initial signs and symptoms	Fuzzy attributes
1. Right ulnar nerve thickening	Marked
2. Single patch over right thumb	Large
3. Loss of sensation	Severe

Interestingly, no new symptoms or signs have been palpable after one month. The evaluator (PERF) compares the beliefs of the respective signs and symptoms in the predicted and observed sets.

*Case II:* There may be some occasions when for some  $s \in \hat{S}_i(D, T^*, P_h, t_0)$  there would be no corresponding  $s$  belonging to  $S_i$ . Similarly, for some  $s \in S_i$  no corresponding  $\hat{s}$  belonging to  $\hat{S}_i(D, T^*, P_h, t_0)$  may be present. These two cases are dealt as follows:

*Case II (a):* For some  $\hat{s} \in \hat{S}_i(D, T^*, P_h, t_0)$  there does not exist any corresponding  $s \in S_i$ . Example: In a case of type I lepra reaction at time  $t_0$  the legs and face of the patient were moderately swollen (oedema: moderate), there was severe aching all over the joints and there was neurological deficit grade II. On treatment with Clofazimine (300 mg) daily, Ibuprofen (400) twice daily, vitamin B-complex and adequate antacid coverage for seven days, the expected set of signs and symptoms  $\hat{S}_7$  becomes:

Oedema: less marked; aching: less acute; and neurological deficit: very low.  
Suppose that after seven days the following observations ( $S_7$ ) are reported:  
Oedema: moderate; and aching: moderate.

Note that the observed set does not include any information about the third symptom. Therefore, there is difficulty in computing a straightforward "match" or "mismatch" by the evaluator. This problem is resolved by the evaluator which recommends information regarding the missing symptom to be gathered. Once necessary information is fed, the matching process is repeated again.

*Case II (b):* For some  $s \in S_i$ ; there does not exist any corresponding  $\hat{s} \in \hat{S}_i(D, T^*, P_h, t_0)$ .

For example, in a case of borderline tubercular leprosy the observations made are reported in table 1.

On treatment with Dapsone and Rifampicin for one month, the expected set  $\hat{S}_{onemonth}$  would be as given in table 2.

**Table 2.** Predicted signs and symptoms and their fuzzy attributes.

Predicted signs and symptoms	Fuzzy attributes
1. Right ulnar nerve thickening	Less marked
2. Patch over right thumb	Moderate
3. Loss of sensation	Moderate

**Table 3.** Actual signs and symptoms and their fuzzy attributes after one month.

Actual signs and symptoms after one month	Fuzzy attributes
1. Right ulnar nerve thickening	More marked
2. Patch over right thumb	Large
3. Loss of sensation	Moderate
4. Tenderness over right ulnar nerve	Marked
5. Aching over the joints	Mild to moderate

On observation after seven days the actual set of signs and symptoms  $S_{onemonth}$  contains elements as listed in table 3.

Here two new signs and symptoms (4) and (5) appear in the set  $S_{onemonth}$  which were altogether absent in  $\hat{S}_{onemonth}$ . The evaluator (PERF) takes some extra care to deal with such cases. Firstly, it computes the similarity measure among the remaining elements of  $\hat{S}_i(D, T^*, P_h, t_0)$  and  $S_i$  and modifies the belief of previous provisional diagnosis in PROG by sending a "match" or "mismatch". Simultaneously, it takes the help of HEXPERT through the module PG1 in PROG (described below) to find another disease  $d_k \in D_{t_0}$  for which the predicted and the observed set of symptoms match. The process is repeated till no disease exists in  $D_{t_0}$  for which the predicted and observed symptoms are found to match.

#### 4.4 PROG

The expert PROG has been divided into three knowledge modules. The first module PG1 interacts with the evaluator module. It gets the input from the evaluator which identifies whether a "match" or "mismatch" has occurred between the predicted and the observed symptoms.

Suppose that the evaluator has sent a match between  $\hat{S}_i(D, T^*, P_h, t_0)$  and  $S_i$  to PG1 which in turn activates PG2. At the same time PG2 also receives the admissible

**Table 4.** Frames and attributes used in LEPDIAG.

List of frames	Example attributes
Frame Control	Status
Complaints	Chronic Ulcer
Checks signs	Regional Anaesthesia
InvestI	Blood Sugar (PP)
InvestII	Skin Smears
PRDI	Tabes Dorsalis
PRDII	Pytirisias Versicolor
PRDIII	Lepra ReactionI
Immune Status	Bacterial Index
Leprosy	Borderline Lepromatous
General Status	Intercurrent Infection
Drug	Clofazimine

set of  $\langle D, T^* \rangle$ , from DIAG in the forward loop at the time instant  $t$ . Since the evaluator has indicated a match between  $\hat{S}_t(D, T^*, P_h, t_0)$  and  $S_t$ , PG2 then increments the belief about the disease  $D$  in the provisional diagnosis  $D_t$  arrived at by DIAG. If  $T^*$  is the treatment recommended by DIAG at time  $t$  for the disease  $D$  i.e.,  $\langle D, T^* \rangle \in \langle D, T \rangle_t$ , the belief about  $T^*$  is also incremented. PG2 then examines the elements of newly generated  $\langle D, T \rangle_t$  with their associated beliefs. Finally, the disease and treatment pair  $(D^*, T^*)$  having highest belief is selected and passed to the next module PG3.

The evaluator sends a "mismatch" to PG1. The module communicates to the homeostatic expert for finding the predicted set of symptoms  $\hat{S}_t(D, T^*, P_h, t_0)$  for an other  $d_k$  in the set  $D_{t_0}$ . This process is repeated till all the elements of  $D_{t_0}$  are exhausted or a single case of "match" is arrived at by the evaluator. If a "match" appears, it is dealt with as described above. But if all the elements of  $D$  are exhausted without finding a "match", then the forward loop predominates and PG2 only accepts  $\langle D, T^* \rangle$  supplied by DIAG at time  $t$ .

The knowledge module PG3 is concerned with the therapeutics of different diseases. The module PG3 receives as input the diagnosed disease-entity and its treatment from PG2. This module also enquires about patient's general status and drug toxicity parameters. It then optimizes the drug elements (along with the adjustment of doses of different drugs for different individuals). For example, if the most possible diagnosis as given by the module PG2 is borderline leprosy and the proposed drug no. 1 is Dapsone then PG3 asks for the following facts:

- (1) whether any tenderness is present over the adjoining peripheral nerve;
- (2) whether there is any past history of lepra reaction present in the patient;
- (3) whether the skin patches are looking considerably inflamed;
- (4) whether the patient is running a temperature;
- (5) percentage of haemoglobin in the blood sample of the patient.

Suppose the replies to the queries for a particular patient are as follows:

- (1) Considerable; (2) history of lepra reaction three years earlier; (3) slight; (4) mild; (5) 10 gm%.

Then the module PG3 indicates that:

- (1) administration of steroid 20 mg daily is indicated;
- (2) oral administration of Clofazimine 300 mg daily to be started;
- (3) oral antacids and antitflatulents are to be administered;
- (4) to report to the expert after one week.

## 5. Knowledge representation and inferencing

The information entities in LEPDIAG are organized in the form of objects. Each such object consists of a number of attributes. These objects and attributes are implemented as frames and slots. A list of typical objects are enumerated in table 1.

LEPDIAG has been built around an expert system building tool FEXT. FEXT (fuzzy expert tool) is an enhanced version of DEXT (Basu *et al* 1988) and incorporates the features required for dealing with fuzzy information and knowledge. In FEXT expert directives and heuristics are encoded through production rules. These domain specific rules are stored in the rulebase in the form of 'condition-action' type of rules. The

rules are represented as

Rule::= IF <antecedent field> THEN <consequent field> <BM>,

where <BM> stands for belief measure, which is a real number or a fuzzy qualifier indicating a heuristic weight or fuzzy truth value associated with a rule. The working memory of FEXT has been organized as a collection of "frames". FEXT views every feature of a problem as an information object which may consist of several component pieces of information.

**Belief measure (BM):** In certain situations, the state of the working memory (WM) may match the antecedent field of more than one rule, resulting in a conflict among the match rules. The conflicting rules often suggest different diagnoses, decision algorithms etc. In such cases, the expert's heuristics can be applied for resolving the conflict. The expert's heuristic preferences are encoded in the form of belief measures associated with the rules. BM represents the confidence or belief of the expert in the truth of a rule. As detailed in § 5.1, BM relates to  $\tau$ , the truth qualifier of a rule. Note that  $\tau$  can be a crisp value or a fuzzy set. Consequently, in FEXT, BM can be coded either as a positive real number or in terms of a fuzzy linguistic qualifier and the composite belief about the multiple facts obey the principle of arithmetic operations on fuzzy numbers (Kandel 1986; Leung & Lam 1988). The implementation of BM is in sharp contrast to the confidence factor (CF) used in MYCIN and others where only real numbers have been used as CF and the composite belief has been computed as a simple multiplication of the CF. The methodology for computing such beliefs has been explained in § 5.1. This measure also is utilized for resolving conflicts in conjunction with meta rules as explained in § 5.2.

FEXT allows the rules to be considered as a number of rule clusters depending on the functional characteristics of the rules. Also special constructs are available in FEXT for intercluster communication and context switching. Thus in the context LEPDIAG, FEXT facilitates the modelling of different knowledge modules as distinct rule clusters.

*Example rules of LEPDIAG:* This rule belongs to rule cluster II, rule no. 7.

IF (\$checksigns # skin patch = flat)  
(\$checksigns # patch pigment = hypo)

Then

Set (\$PRDII # vitiligo = considerable)  
AND Set (\$PRDII # pityriasis-versicolor = likely)  
AND Set (\$PRDI # leprosy = more likely).

The rule should be read as if the observed skin patch over the patient's body is flat and hypopigmented then the belief about the diagnoses of vitiligo, pityriasis-versicolor and leprosy would be 'considerable', 'likely' and 'more likely' respectively.

The inference machine of FEXT adopts forward chaining of inferencing. Also, in tune with the medical domain, the knowledge is often imprecise and fuzzy. The fuzzy inferencing required by LEPDIAG is captured by FEXT as discussed below.

### 5.1 Fuzzy inferencing

Definitions and concepts about the fuzzy set theory are available in the literature (Kaufman 1975). In the fuzzy set literature, a rule If  $X$  is  $A$  then  $Y$  is  $B$  is treated

as a conditional fuzzy proposition which defines a fuzzy relation over  $U \times V$ . There are several alternative translation schemes for defining the fuzzy relation associated with a conditional fuzzy proposition (Zadeh 1983b).

In LEPDIAG, we have used the translation rule  $R_g$  defined below to determine the possibility distribution associated with a conditional fuzzy proposition.

# DEFINITION

Let  $F$  and  $G$  be fuzzy subsets of  $U$  and  $V$ , respectively. The possibility distribution  $\pi(X \rightarrow Y)$  associated with the conditional fuzzy proposition **If  $X$  is  $F$  then  $Y$  is  $G$**  is given by

$$\pi(X \rightarrow Y) = R_g \quad (1)$$

where  $R_g$  is a fuzzy subset of  $U \times V$  with membership function

$$\mu_{R_g}(u, v) = 1, \text{ if } \mu_F(u) < \mu_G(v) = \mu_G(v), \text{ otherwise.}$$

Suppose that  $A$  and  $B$  are fuzzy sets over the universes  $U$  and  $V$  respectively. Consider a fuzzy rule: **If  $X$  is  $A$  then  $Y$  is  $B$** . Suppose that one intends to use this rule with a fact  $X$  is  $A1$  where  $A1$  is a fuzzy set over  $U$ . By generalised *modus ponens* (Nilsson 1982), one can infer:

*Ant1*: If  $X$  is  $A$  then  $Y$  is  $B$

*Ant2*:  $X$  is  $A1$

*Cons*:  $Y$  is  $B1$

where  $B1$  is a fuzzy set over  $V$ . The possibility  $B1$  of  $Y$  is computed using the translation rule  $R_g$  for conditional fuzzy proposition and the compositional rule of inference (Mizumoto and Zimmerman 1982). Accordingly, the membership function for  $B1$  is given by:

$$\mu_{B1}(y) = \max_x(\min(\mu_{A1}(x), \mu_{R_g}(x, y))).$$

When the rule is truth qualified such as if  $X$  is  $A$  then  $Y$  is  $B$  is  $\tau$ , where  $\tau$  is a fuzzy set over  $[0, 1]$ , then with a fact  $X$  is  $A1$  the possibility distribution  $B1$  of the consequent is obtained by applying the translation rule for truth qualified propositions and compositional rule of inference. Thus we have:

If  $X$  is  $A$  then  $Y$  is  $B$  is  $\tau$  Rule

$X$  is  $A1$  Fact

---

$Y$  is  $B1$  Consequent

The membership function of  $B1$  is given by:

$\mu_{B1}(y) = (\mu_\tau)_x(\max(\min(\mu_{A1}(x), \mu_{R_g}(x, y))))$ . We can generalize the inference procedure to the cases where a rule involves more than one antecedent. For instance, consider a rule:

If  $X_1$  is  $A_1$  and  $X_2$  is  $A_2$  and, ...,  $X_k$  is  $A_k$  then  $Y$  is  $B$ .

In this case, the translation rule defines a fuzzy relation  $R_g$  over  $U_1 \times U_2, \dots, U_k \times V$  where  $U_i$  is the universe for  $X_i, i = 1, \dots, k$ , and  $V$  is the universe of  $Y$ . The membership function for  $R_g$  is given by:

$$\mu_{R_g}(x_1, x_2, \dots, x_k, y) = 1, \text{ if } \min(\mu_{A1}(x_1), \mu_{A2}(x_2), \dots, \mu_{AK}(x_k)) \leq \mu_B(y), \\ = \mu_B(y) \text{ otherwise.}$$

This follows from the translation rules for combining two fuzzy propositions using AND.

Suppose now this rule is used with facts:  $X1$  is  $A1'$ ,  $X2$  is  $A2'$ , ...,  $Xk$  is  $Ak'$ , the possibility distribution  $B1$  of the consequent  $Y$  is then given by:

$$\mu_{B1}(y) = \max(\min(\min(\mu_{A1'}(x_1), \dots, \mu_{Ak'}(x_k)), \mu_{R_g}(x_1, \dots, x_k, y))).$$

## 5.2 Problems in inferencing

**Rule conflicts:** When the rule base contains more than one rule having the same consequent (say  $Y$ ), then based on a given set of facts these rules may lead to different beliefs about  $Y$ . For instance, consider the rules:

Rule  $r_1$ :  $X1$  is  $A1$  and  $X2$  is  $A2$  then  $Y$  is  $B$

Rule  $r_2$ :  $X1$  is  $A1$  and  $X3$  is  $A3$  then  $Y$  is  $C$ .

Suppose that at some stage of inferencing the facts are  $X1$  is  $A1'$ ,  $X2$  is  $A2'$  and  $X3$  is  $A3'$ . Then based on  $r_1$  and  $r_2$  we would have  $Y$  is  $B1$  and  $Y$  is  $C1$ , respectively, where the fuzzy beliefs  $B1$  and  $C1$  about  $Y$  can be computed following the procedure described in the preceding section.

It may be mentioned that two rules may lead to totally conflicting beliefs about  $Y$ , e.g., when  $B = \bar{C}$ . For instance, the following two rules:

Rule#41 in Block \$2.

If  
Skin patch is considerably raised and moderately pigmented  
AND history of primary syphilis is almost definite  
Then  
Acquired syphilis is likely.

Rule#33 is Block \$4.

If  
Peripheral nerve thickening is moderate  
AND VDRL titre is weakly reactive  
Then  
Acquired syphilis is unlikely.

Suppose that a patient exhibit all the four symptoms to a certain extent, then the two rules would lead to contradictory beliefs about the diagnosis "acquired syphilis".

It is therefore necessary to adopt a suitable scheme for estimating the final belief about the consequent based on conflicting rules. In fuzzy set literature (Kandel 1986) several schemes have been suggested for estimating the final belief about the consequent. But usage of any particular scheme to deal with all types of rule conflicts often lead to unsatisfactory solutions. For example, one approach may be to select  $Y$  to  $\hat{B}$  where

$$\hat{B} = B \cap C.$$

This scheme would unnecessarily favour the negative evidence. Such an approach would be unacceptable in a situation where belief about the rule generating negative

evidence is low. Alternatively, if we select  $\hat{B} = B \cup C$ , any negative evidence will have lower weightage.

In view of this, LEPDIAG uses meta rules to combine the beliefs. A typical meta rule may be:

$MR_1$ : If two rules provide conflicting evidence about a consequent, then select the rule having higher belief.

The conflict resolution strategy followed in LEPDIAG does not use a single meta rule to deal with all types of rule conflicts. Rather, depending upon the context, different meta rules are used. For instance, instead of the meta rule  $MR_1$ , LEPDIAG may sometimes use the following metarule:

Let  $R_1$  and  $R_2$  denote two sets of rules having same consequent  $Y$ , where the rules in  $R_1$  provide positive evidences about  $Y$  and the rules in  $R_2$  provide negative evidences (e.g.,  $Y$  is unlikely). Suppose that based on the observed facts (e.g. symptoms and signs) the rules in  $R_1$  and  $R_2$  are fired. Suppose a rule  $r_{i_1} \in R_1$  where  $i = 1, \dots, k$ , generates a belief  $Y$  is  $B_i$  and a rule  $r_{j_2} \in R_2$ ,  $j = 1, \dots, m$  provides a belief  $Y$  is  $C_j$ .

The meta rule ( $MR_2$ ) now computes the final belief  $Y$  is  $A$ , where

$$A = \bigcup_{i=1}^k B_i \odot \bigcap_{j=1}^m C_j,$$

and where the binary operator  $\odot$  estimates the membership value as the average of the membership values in the operand fuzzy sets. Thus

$$\mu_A(\tau) = \frac{1}{2}(\mu_B(\tau) + \mu_C(\tau)),$$

where,

$$B = \bigcup_{i=1}^k B_i$$

and

$$C = \bigcap_{j=1}^m C_j.$$

As explained in the next section, the expert system shell used to build LEPDIAG partitions the rule base into different clusters depending on suitable criteria. LEPDIAG utilises appropriate meta rules to resolve the rule conflicts among the rules in a cluster.

**5.2a Problem of incomplete evidences about antecedents:** Consider a rule ( $r$ ):

If  $V_1$  is  $A_1$  and  $V_2$  is  $A_2$  and ...  $V_n$  is  $A_n$  then  $D$  is  $B$  is  $\tau$ , where  $V = v_1, v_2, \dots, v_n$  is a set of symptoms and signs.

Suppose a patient arrives with asymptoms and signs  $V_0 = v_j, \dots, v_m$  where  $V_0 \in V$ . Thus the values of all the signs and symptoms required for the rule  $r$  are not known at this stage. In medical diagnosis it often becomes necessary to find the possibility of the disease  $D$  based on available symptoms and signs only. This may be done with the help of the projection principle which reduces the rule  $r$  to  $\hat{r}$ .

$$\hat{r}: \text{If } V_j \text{ is } A_j \text{ and } \dots, V_m \text{ is } A_m, \text{ then } D \text{ is } B_i.$$

The use of the projection principle, however, introduces a few problems. First, too many rules may be fired on some meagre quantum of evidences making search and computation time larger than expected. Second, different projections may generate different beliefs about the consequents. In view of this, LEPDIAG uses meta rules to deal with the problem of incomplete evidence.

To illustrate a typical meta rule used in LEPDIAG, consider a set of conflicting rules  $r_1, \dots, r_k$  where each  $r_i$  determines the belief about the same disease  $D$ . Let  $W_i$  denote the set of symptoms and signs which appear in the antecedent of rule  $r_i$ . Suppose that observed signs and symptoms are  $V$ , where  $V \subseteq W_i$ ,  $i = 1, \dots, k$ . Then a meta rule in LEPDIAG selects the rule  $r_j$  such that  $\text{card}(W_j \cap V)$  is maximum, i.e., the rule having maximum information about its antecedents is selected. The projection principle is used to reduce the rule  $r_j$  to  $\hat{r}_j$  such that it involves only the signs and symptoms in  $V$ . The reduced rule  $\hat{r}_j$  is now fired to compute the belief about  $D$  from the values of observed signs and symptoms. If more than one rule qualifies the selection criterion e.g., when  $\text{card}(W_j \cap V) = \text{card}(W_p \cap V)$ , then all these rules are selected while computing the belief about  $D$ . The final belief about  $D$  based on such conflicting rules is computed following the procedure described in the previous section.

### 5.3 Belief modification

As described in the previous section, the module PERF examines the degree of match between the observed and the predicted (by the disease model) signs and symptoms and to transfer the match-information to the next knowledge module PROG. PROG finally changes the belief about the proposed diagnoses and treatment.

Since the symptoms and signs can assume fuzzy values, the comparison between predicted and observed symptoms is also carried out using fuzzy logic. PROG receives match or no match from PERF with respect to  $S_i$  and  $\hat{S}_i(D T^*, P_h, t_0)$ . Based on the degree of match between the observed and predicted symptoms and signs reported by PERF, the module PROG modifies the belief about the current diagnosis obtained by DIAG. Again meta rules are used to carry out such belief modification of the final diagnosis. Before proceeding to describe the belief modifications in PROG, the methodology followed by PERF to compute the degree of match is briefly described here.

*Match:* Suppose that for a symptom or a sign  $X$ , the observed and predicted values are,

Given  $X$  is  $A_1$  in  $S_i$ ;  
and  $X$  is  $A_1'$  in  $\hat{S}_i(D T^*, P_h, t_0)$ , respectively.

To decide whether these two observations match, we need to define suitable fuzzy resemblance relations 'match' among the set of fuzzy sets ( $F$ ) over the universe of  $X$ . The relation match should satisfy the following properties.

*Reflexivity:*  $\text{match}(A, A) = 1.0 \ A \in F$

*Similarity:*  $\text{match}(A, B) = \text{match}(B, A) \ A, B \in F$ .

A typical match relation used in LEPDIAG is given below:

i) Both  $A$  and  $B$  are crisp values i.e.,  $X$ -values are precisely known.

$\text{match}(A, B) = 1.0$  if  $A = B$

$= 0.0$ , otherwise.

ii) Suppose  $A = a$  is crisp and  $B$  is a fuzzy set

$\text{match}(A, B) = \mu_B(a)$

Alternatively, when  $A$  is a fuzzy set and  $B = b$  is crisp,

then

$\text{match}(A, B) = \mu_A(b)$



iii) Both  $A$  and  $B$  are fuzzy sets.

$$\text{match}(A, B) = \max((\text{card}(A \cap B)/\text{card}(A)), (\text{card}(A \cap B)/\text{card}(B))),$$

where  $\text{card}(A)$  denotes the cardinality (Kandel 1986) of the fuzzy set  $A$ .

The resemblance relation  $\text{match}$  can be readily extended over a set of symptoms and signs.

Let  $S_i = X_1$  is  $A_1, X_2$  is  $A_2, \dots, X_k$  is  $A_k$

$\hat{S}_i(D, T^*, P_h, t_0) = X_1$  is  $B_1, X_2$  is  $B_2, \dots, X_k$  is  $B_k$

Then

$$\text{match}(S_i, \hat{S}_i(D, T^*, P_h, t_0)) = \min(\text{match}(A_1, B_1), \text{match}$$

$$(A_2, B_2), \dots, \text{match}(A_k, B_k)).$$

PERF uses a suitable  $\alpha$  cut off to decide whether observed and predicted signs and symptoms do match or not. Thus  $S_i$  and  $\hat{S}_i$  are said to match if,

$$\text{match}(S_i, \hat{S}_i) > \alpha, \text{ else PERF informs no match.}$$

( $\alpha$  - Cut): The choice of  $\alpha$  - cut is highly system-dependent. If the homeostatic model contains satisfactory statistical information about the future genesis of signs and symptoms for a given disease, then the cut-off value of  $\alpha$  can be assumed to be closer to 1.0 and consequently there would be a tighter match between the observed and the predicted signs and symptoms.

On the other hand, unsatisfactory information content of the same model requires a less tight match and the value of  $\alpha$  would be closer to 0.0. For our convenience, we have set the value of  $\alpha$  as 0.5 and with trials of further experimental data attempt to push the value towards 1.0.

**5.3a Belief modification by PROG:** Based on the degree of match between the observed and predicted symptoms reported by PERF, the expert system PROG modifies the belief about the current diagnosis with the help of meta rules. To illustrate the steps taken by PROG,

let belief about  $D_j$  and  $T^*$  obtained by DIAG based on  $S_i$  be:

$D_j$  is  $\tau_j$  and  $T^*$  is  $\alpha_j$ .

In case of a match, a typical PROG metarule may modify the beliefs to:

$D_j$  is  $f_{\text{match}}\tau_j$  and  $T^*$  is  $g_{\text{match}}\alpha_j$ .

On the other hand for no match (i.e.,  $\mu_{\text{match}}(S_i, \hat{S}_i) < \alpha$ ) the beliefs are modified to:

$D_j$  is  $f_{\text{nomatch}}\tau_j$  and  $T^*$  is  $g_{\text{nomatch}}\alpha_j$ .

where typical  $f_{\text{match}}$  and  $f_{\text{nomatch}}$  may be "very" and "more or less" respectively.

LEPDIAG utilises different meta rules on different occasions to change the beliefs about diagnoses and treatment. Thus another meta rule may modify the beliefs as follows:

Suppose PERF returns:

$$\text{match}(S_i, \hat{S}_i(D_j, T^*, P_h, t_0)) = \beta.$$

The PROG meta rule MRPROG-2 modifies the belief about  $D_j$  to  $\beta * \tau_j$ , where  $*$  is a fuzzy arithmetic operator (Kandel 1986).

## 6. A case study

### 6.1 Consultation with DIAG

Suppose a patient arrives at the clinic with the complaint of a 'solitary' white patch on his right leg and "considerable" loss of sensation over the skin of the patch area. Based on these observations, DIAG selects a set of candidate diseases as provisional diagnoses. Each disease among the set is also assigned with some approximate belief about their occurrence. In this case, three different groups of disease entities containing 38 diseases are thus identified.

- (1) Prov-diag-nerves – peripheral neuropathy, tabes dorsalis etc. (likely)
- (2) Prov-diag-skin – scleroderma, myxoedema, pityriasis versicolor etc. (likely).
- (3) Prov-diag-more – type I allergic disorders, nephrotic syndrome etc. (likely).

For further refinement of diagnosis, the patient is examined for elicitation of more facts in the same context. Though a generalised search is done using anatomical, physiological and pathological heuristics, provisional diagnoses are done based on evidence of whether some specific cardinal features are present. Here, DIAG enquires about the following signs:

S1. Nerve thickening; S2. regional anaesthesia; S3. muscle wasting; S4. flat and hypopigmented patch; S5. raised and pigmented patch; S6. generalised thickening of skin.

Suppose in the present case, there is evidence of "doubtful" right musculocutaneous nerve thickening and "diffuse" regional anaesthesia over the skin of the right leg. These pieces of information select only a subset of provisional diagnoses out of a list of 38 diseases. The possibilities of the provisional diagnoses are estimated again using the fuzzy rule bases. Consequently, the following diseases have "more likely" possibilities:

*PRD1*, leprosy; *PRD2*, Peripheral neuropathy; *PRD3*, Syringomyelia; *PRD4*, tabes dorsalis; *PRD5*, Hereditary sensory radicular neuropathy; *PRD6*, congenital indifference to pain; *PRD7*, hysteria.

In this test case, suppose there was no evidence of the abnormalities S1 and S2 present since birth. Consequently DIAG reduces the belief measure of *PRD5* and *PRD6* to less likely. Again, there was no evidence of objective sensory loss, hence the possibility level of *PRD7* is reduced to "never" indicating its strong negative confirmability. Further, the patient was found not to bear any evidence of dissociated sensory loss, i.e., pinprick and thermal sensation is lost, but touch sensation is spared (Harrison 1988, p. 675). Therefore the belief of *PRD3* is reduced to "almost never". At this stage the patient was tested for the following three cardinally important signs to prove or disprove *PRD2*.

- (1) Difficulty in walking; (2) loss of position sense; (3) broad based stamping gait.

But the presence or absence of these signs could not be ascertained satisfactorily.

Therefore to resolve the dilemma a set of special investigation reports becomes very essential.

## 6.2 Directives for special investigations

In order of preference (according to cost and invasiveness of the special investigations) DIAG recommends the following investigations.

I0. Blood sugar (fasting, F); I1. Blood sugar (*post-prandial*, PP); I2. Wasserman Reaction; I3. Kahn Test; I4. VDRL Test.

Suppose now, with the patient under study, the results of the above mentioned investigations are as follows:

Fasting blood sugar level is 80 mg; Wasserman Reaction and Kahn test are negative, VDRL test is weakly positive.

Unfortunately, these investigation results are within the range of their respective normal values. Hence the system confirms *PRD2* and *PRD4* to be "impossible". With this, the only class of provisional diagnosis left for consideration is leprosy, to be investigated further. On further investigation, the lepromin test returns 1+ and skin smear shows no acid fast organism. As the patient has a "solitary" hypopigmented patch, "diffuse" regional anaesthesia of right leg, Lepromin test result 2+ enhances the confidence of the *PRD* leprosy to the level of "almost always".

## 6.3 Type and status of the disease-entity

As soon as the diagnosis leprosy is arrived at, DIAG recommends much closer examination of the patient to map the patient on the logical states of the full spectrum of the disease. There are five clinical and nine immunohistological reference states to be considered. For clinical establishment of the class of leprosy from which the patient is suffering, a battery of clinical evidence is sought for. This includes –

- (E1) diffuse thickening of skin (frank, ..., absent);
- (E2) number of skin patches or macules (nil, few, ..., innumerable);
- (E3) nodules present over the patch (distinct, ..., absent);
- (E4) polymorphic "punched out" centre (distinct, ..., absent);
- (E5) irregular plaques of skin (absent, few, ..., definite);
- (E6) partially raised edges of skin (absent, ..., definite);
- (E7) satellite patches (specific, less specific, ..., absent);
- (E8) distribution of the patch (bilaterally symmetrical, irregular symmetry, ..., completely asymmetrical);
- (E9) size of the macule (very small, fairly small, small, ..., large);
- (E10) clarity of the margin (absolute, notable, unspecific, ..., nil).

In the present case, suppose there are very few skin patches, definitive raised skin margins around the patch, satellite patches numbering a few and notable clarity of the margin of the lesion. These observations, especially responses from (E2), (E6), (E7) and (E10) enhance the possibility of clinical type *borderline tubercular leprosy* to the level "confirmed". At the same time, the negative evidence of (E1), (E3)–(E5) and (E8)–(E9) disproves the other clinical types of leprosy.

#### 6.4 Consultation through HEXPERT

To pinpoint the patient on the internal disease spectrum a set of investigative parameter values are sought for. These are: I1. Lepromin Test, I2. skin biopsy, I3. nasal smear, I4. leucocyte migration inhibition test, I5. lymphocyte transformation test, I6. fluorescent antibody test, I7. nerve biopsy, I8. drugs administered, I9. sample time.

With the patient under investigation, the response for I1 = 2 + ; I2 = "large number of" lymphocytes found; I3 = *M. leprae* less than 5%; I4 = 76%, I5 = "more than" 6%; I6 = "very low" titre; I7 = infiltration of immunocytes "huge".

The drugs chosen by PROG are Dapsone and haematinics. These drugs are administered to the patient for one month. The system now uses the above values in the model of leprosy-immune-reaction and predicts the expected signs and symptoms of the same patient after a one-month interval. The predictions are as follows:

- P1: size of the patch would be smaller;
- P2: area of anaesthesia would be less diffuse;
- P3: satellite lesions would be less marked;
- P4: development of areas of new patches would be unlikely.

#### 6.5 Consultation through PERF

After one month the following observations about the patient have been reported:

- O1: size of skin patch is increased;
- O2: area of anaesthesia has become more diffuse;
- O3: satellite lesions are more distinct;
- O4: fresh crops of skin patches appear on the left leg.

Note that there is serious discrepancy between the prediction of HEXPERT and the observed symptoms. The system "watchman" PERF, declares a "mismatch" among the predicted and the observed signs and symptoms. The message is conveyed to PG1 which in turn directs HX1 to generate signs and symptoms for other elements in (D, T). HX1 now selects the diagnosis Dapsone-resistant Borderline Tubercular Leprosy as the next candidate and generates the signs and symptoms to be used by PERF. The new choice results in a better match with the observed signs and symptoms. Accordingly, PERF invokes PG1 to suggest therapeutic advice to the patient.

#### 6.6 Consultation through PROG

Before proceeding towards therapeutic advices, LEPDIAG asks some further questions about the general status of health of the patient which is very much essential for optimisation of the treatment regimen. Necessary information about the general status includes the following.

- (1) anaemia; (2) jaundice; (3) history of jaundice within past two years;
- (4) known allergy to sulphur; (5) amount of urine passed in last 24 h;
- (6) difficulty in urination, (7) colour of the urine.

Suppose, the patient under investigation has only a history of jaundice one year

earlier. Therefore, the drug Rifampicin is to be avoided in this particular case because the drug is a known hepatotoxic agent. Therefore, PROG finally prescribes the following set of drugs for the patient:

- (1) Clofazimine, (2) vitamin B-complex, (3) BCG vaccine as an immunostimulant.

A new sample time is set for the patient and the same procedures would be repeated till the patient is declared "cured" of the disease leprosy. Table 4 summarises the frames and attributes used in LEPDIAG.

## 7. Other issues

No expert system implementation is feasible without building a powerful knowledge-base. The vital task of knowledge acquisition was simplified in the case of LEPDIAG, through the active participation of a medical practitioner (the first author) all through the development phase. Instead of using any computer tool, as used in TEIRESIAS (Davis 1978, pp. 99–134), we have formed the knowledge-base through intense discussions and interviews. The knowledge thus compiled, was then structured into different functional categories. This phase went through a number of refinement iterations as the architecture of LEPDIAG was being incrementally refined. The knowledge was then coded into a fuzzy production rules form with the fuzzy beliefs being attributed by the doctor. The knowledge-base was built into LEPDIAG through the use of the expert system building tool FEXT. FEXT supports the knowledge-edition feature, where the rules are parsed, compiled and internally stored as strings of tokens.

The system has been validated by applying the system on a number of case histories fed as inputs and comparing the outcome of the system with those resulting from expert medical practitioners. The case histories and the treatment plans followed, were obtained from the archives of the Department of Leprology, School of Tropical Medicine, Calcutta. We obtained reports of 240 cases over a period of three years. About 120 rules have been developed over 400 slots (attributes) distributed over 17 frames (objects). We experimented by feeding information of 10 cases. We simulated the actual user response and examined the response of LEPDIAG. A comparison of LEPDIAG responses with actual doctors prescriptions was found to be tallying for about 60% of the cases. For the rest, LEPDIAG responses were less accurate as some special combination of symptoms were not supported by the existing LEPDIAG rules. We have planned to adopt a deep-reasoning technique to alleviate this limitation.

The present version of LEPDIAG does not support automated testing of knowledge consistency and completeness. However, for static evaluation of rulebase consistency, we are working on modelling the rule base as Petri nets and are in the process of formulating some consistency criterion. Recent works by other researchers have also proposed the use of Petri nets for this purpose (Murata *et al* 1991).

LEPDIAG supports explanations in the forms of "how" and "why" queries. For the first query the system explains how a particular slot has been instantiated to a particular value. The second query is relevant when the system asks for some user input. In response to this query the system displays the textual version of the rule prompting the seeking of user response. Presently, explanations are being generated by a rule tracing and instance stack traversal technique and are thus of limited capacity. Incorporation of deep reasoning and storage of deeper models will make the explanation generation more meaningful.

## 8. Conclusion

A knowledge-based system capable of long-term monitoring of chronic endemic diseases has been proposed in this paper. Multiple experts have been inducted for the desired purpose. Effective cooperation among different expert systems within the parent body enables the composite system to offer more efficient therapies for the changed disease-states over time. A homeostatic model keeps pace with the changed environment of the patient's body. Even a mistaken diagnosis and a mistaken treatment regime can be corrected by the introduction of PERF – the performance evaluator. Moreover, the system counts the general status of health for each patient in each session before prescribing the final set of treatments to be administered.

The domain knowledge has been organised as partitioned fuzzy production rules. The paper discusses the knowledge representation alongwith "Meta rule-based conflict resolution" techniques adopted for resolving fuzzy decision conflicts.

This approach can be utilized for any chronic disease process for on-line diagnosis, monitoring prognosis and more efficient management.

## Glossary of medical terms

**Homeostasis:** The capability of a system to hold its Critical Variables within physiological limits in the face of unexpected disturbance or perturbation.

**Idiotypic network:** The collection of determinants on a particular antibody, that is recognised by other antibodies, is called its idio type, and the antibodies performing the recognition are said to be antiidiotypic. All antibodies have idiotypic determinants, and every antibody should be able to recognise at least some idio type. Thus the immune system forms a large interconnected idiotypic network.

**Immunological reaction:** The immune system in vertebrates consists of nearly  $10^8$ – $10^{12}$  cells. The cells are usually of two main types: *B*-lymphocytes and *T*-lymphocytes. The *B*-cells secrete antibodies which bind to and hasten the elimination of foreign antigen. *T*-cells lyse the foreign cells alone or with macrophases and also amplify the *B*-cell response. In both kinds of immune reaction it is usually ensured that the response does not become self-destructive.

**Prognosis:** During any disease process in any person, patient may be driven to either complete cure or towards more debility upto death. The term prognosis is meant to gauge the direction of the disease process at any point of time. A good prognosis points to cure and bad one towards debility.

## References

- Adlassnig K P, Kolarz G 1982 CADIAG-2: Computer-assisted medical diagnosis using fuzzy subsets. In *Approximate reasoning in decision analysis* (eds) M M Gupta, E Sanchez (Amsterdam: North Holland)
- Albin M A 1975 *Fuzzy sets and their application to medical diagnosis and pattern recognition*. Ph D dissertation, University of California, Berkeley
- Basu A, Majumder A K, Sinha S 1988 An expert system approach to control system. *IEEE Trans. Syst., Man Cybern.* 28: 685–694

- Buchanan B G, Shortliffe E H (ed.) 1984 *Rule based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project* (Reading, MA: Addison Wesley)
- Chandrasekharan B, Gomez F, Smith J 1979 An approach to medical diagnosis based on conceptual structures. *Proc. Sixth Int. Joint. Conf. Artif. Intell.* (California: Morgan Kaufman)
- Chilauksy R, Jacobsen B, Michalsky R S 1976 An application of variable valued logic to inductive learning of plant disease diagnostic rules. *Proc. Sixth Annual Symposium on Multivalued Logic*
- Cios K J, Shin I, Goodenday L S 1991 Using fuzzy sets to diagnose coronary artery stenosis. *IEEE Trans. Comput.* 24: 57-63
- Clancey W J 1983 The epistemology of a rule-based expert system: a framework for explanation. *Artif. Intell.* 20: 215-251
- Davis R 1978 Knowledge acquisition in rule-based systems: knowledge about representations as a basis for system construction and maintenance. In *Pattern-directed inference systems* (eds) D A Waterman, F Hayes-Roth (New York: Academic Press)
- Davis R 1983 Diagnosis via causal reasoning: paths of interaction and locality principle. *Proc. Am. Assoc. Artif. Intell.* 83 pp. 88-94
- Davis R, Buchanan B G, Shortliffe E H 1977 Production rules as a representation for knowledge based consultation programme. *Artif. Intell.* 8: 15-45
- Ganong W F 1989 *Review of medical physiology* (Englewood Cliffs, NJ: Prentice-Hall)
- Harrison J 1988 *Principles of internal medicine*, 11th edn (New York: John Wiley)
- Hart P E, Duda R O, Einaudi M T 1978 A computer based consultation system for mineral exploration. Tech. Report, SRI International
- Herzenberg L A, Black S J 1980 Regulatory circuits and antibody responses. *Eur. J. Immunol.* 10: 1-11
- Jerne N K 1973 The immune system. *Sci. Am.* 229/1: 52-60
- Kandel A 1986 *Fuzzy mathematical techniques with applications* (Reading, MA: Addison Wesley)
- Kaufman A 1975 *An introduction to the theory of fuzzy subsets* (New York: Academic Press) vol. 1
- Leung K S, Lam W 1988 Fuzzy concepts in expert system. *IEEE Trans. Comput.* 21: 43-56
- Martin N 1977 Knowledge-base management for experiment planning in molecular genetics. *Proc. Fifth Int. Joint. Conf. Artif. Intell.* (California: Morgan Kaufman) pp. 882-887
- McDermott J, Steele B 1981 Extending a knowledge-base system to deal with ad-hoc constraints. *Proc. Seventh Int. Joint Conf. Artif. Intell.* pp. 824-828
- Mitzumoto M, Zimmerman H J 1982 Comparison of fuzzy reseasoning methods. *Fuzzy Sets Syst.*, 8: 253-283
- Murata T, Subrahmanian V S, Wakayama T 1991 A Petri net model for reasoning in the presence of inconsistency. *IEEE Trans. Data Knowledge Eng.* 3: 281-292
- Nath I 1983 Immunology of human leprosy - current status. *Leprosy Rev.* (Special Issue) 31S-45S
- Nilsson N J 1982 *Principles of artificial intelligence* (Tioga Publication)
- Patil R S 1987 *A case study of evolution of system building expertise: Medical diagnosis in AI in 1980s and beyond* (ed.) W E L Grimson, R S Patil, (Cambridge, MA: MIT Press)
- Perelson A S 1989 Immune network theory. *Immunol. Rev.* 110: 5-32
- Perez-Ojeda A 1976 *Medical knowledge network: A database for computer-aided diagnosis*, thesis, University of Toronto, Toronto
- Rees R J W 1969 New prospects for the study of leprosy in the laboratory. *Bull. World Health Organisation* 40: 785-800
- Reggia J A, Tuhim S (eds) 1985 *Computer assisted medical decision making* (Berlin: Springer Verlag)
- Rich W 1988 *Essential immunology*, 2nd edn (London: Oxford Press)
- Richter P H 1978 *Complexity and regulation of the immune system: The network approach in systems theory in immunology* (eds) C Bruni, G Doria, G Koch, R Strom (New York: Springer-Verlag)
- Ridley D S 1972 Review of the five group system for the classification of leprosy according to immunity. *Int. J. Leprosy* 40: 102-103
- Shafer G 1976 *A mathematical theory of evidence* (Princeton, NJ: University Press)

- Stefik M, Aikins J, Balzer R, Benoit J, Birnbaum L, Hayes-Roth F, Sacordoti E 1982 The organisation of expert system: A prescriptive tutorial. Tech. Report, Xerox Palo Alto Research Center
- Sticklen J 1987 *MDX2: An integrated diagnostic system*. Ph D dissertation, Department of Computer & Information Science, The Ohio State University
- Zadeh L A 1983a The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets Syst.* 11: 199-227
- Zadeh L A 1983b A computational approach to fuzzy quantifiers in natural languages. *Comput. Math. Appl.* 9: 149-184



76

L  
B

## Legal counselling system

M<sup>1</sup> SHASHI, K V S V N RAJU<sup>1</sup> and A LAKSHMINATH<sup>2</sup>

<sup>1</sup>Department of Computer Science & Systems Engineering, and

<sup>2</sup>Department of Law, Andhra University, Visakhapatnam 530 003, India

**Abstract.** Legal reasoning involves case analysis in statutory as well as real world perspectives. The impact of real world perspective on case analysis poses a serious challenge to knowledge engineers for building legal expert systems. A legal expert system intends to provide intelligent support to legal professionals. The proposed legal predictive system is an attempt to predict the most probable outcome of a case according to statutory as well as real world knowledge of the legal domain. The system accepts the current fact situation of a case and analyses it interactively with legal personnel. This work introduces a frame-like knowledge structure, LATTICE, with two-dimensional attributes. This paper contains a detailed discussion on artificial intelligence-based case analysis of theft cases in a real world perspective.

**Keywords.** Legal reasoning; artificial intelligence; legal predictive system; legal expert systems.

### 1. Introduction

One of the basic principles of justice is that 'Justice delayed is justice denied'. It is from this that the Supreme Court of India has carved out the fundamental right to speedier trial from article 21 of the Constitution of India. The present adjudication process requires transformation in view of the high cost of legal services, baffling complications in existing procedures and frustrating delays in securing justice. Formal adjudication should be more of a last resort than it has been in the past. In recent times, efforts have been made to develop alternate adjudication models in the form of *Lok Adalats*, *Nyaya Panchayats* etc. In this context, the authors feel that alternate adjudication machinery can be augmented with modern computers for a greater extent of openness and accessibility thus lending credibility to the dependence of both government and people on these modes of alternate adjudication machinery.

Automation in the legal world was first proposed (Mehl 1958, pp. 755–79) at an International Symposium on "Mechanisation of Thought Processes" held at the National Physical Laboratory in Teddington, London. Law machines were classified by him into two types: documentary machines and consultation machines. Documentary machines are meant for legal information retrieval operations such as storing/retrieving legal provisions and supporting as well as opposing precedents relevant to the given case. A program FLITE (Finding Legal Information Through Electronics),

was developed in 1964 as the earliest full text retrieval system for the US Air Force. LEXIS and WESTLAW (Hafner 1987, pp. 35–42) are some of the recent commercial systems offering interactive retrieval through terminals at the customer's office. Intelligent support cannot be provided for the user while retrieving the precedents owing to the text matching (keyword search) technique followed in these systems. Hafner (1987, pp. 35–42) proposed an AI-based conceptual retrieval system using individual case frames so that search for relevancy can be made based on a concept of the case rather than text matching of certain keywords. Considerable research work has thus been carried out and significant developments have taken place in the area of documentary machines.

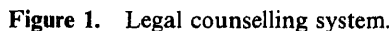
However, no such significant progress can be claimed to have been made in the area of consultation machines which are meant for giving legal advice. The HYPO system developed by Rissland and Ashley (Ashley & Rissland 1988; Ashley 1991) during the 1980s aims at helping an attorney to analyse a new case in the light of relevant precedents (refer appendix A for legal terms), and accordingly generate outlines of arguments for both plaintiff and defendant. The JUDGE system, developed in the late 80s by Bain (Srivastava 1991) proposed modelling the sentencing ability of judges. This system identifies the most similar precedent according to a set of salient features and suggests a commensurate sentence for being awarded in the case on hand. These two systems have been the most widely accepted legal consultation systems to date. But these and similar other consultation systems are oriented towards precedents and are based on a case-based reasoning paradigm.

A precedent can either suggest judgement appropriate to cases with similar current fact situation or it can point to an apt case-law to solve a particular technical ambiguity. These two aspects of the precedent are to be dealt with separately since the first aspect provides only the guidelines whereas the second provides the case-law that is binding on lower courts. The first aspect is emphasised in systems like HYPO whereas the second aspect is considered in systems like JURIX (Srivastava 1991) and Gardner's legal reasoning system (Gardner 1987). Gardner's approach suggests that the case be analysed keeping in view statute as well as relevant case-laws. This system aims at giving decisions for 'easy' cases, while the 'hard' cases, cases which can be argued in either way by a competent lawyer, are left undecided. McCarthy's TAXMAN project (McCarthy 1980) models deductive legal reasoning based on statute. The control strategy of legal systems determines the applicability of those systems to various fields of legal domain – HYPO suits trade secret misappropriation. TAXMAN models the taxation of corporate reorganisation. Gardner's system deals with formation of contracts by offer and acceptance. However, for certain other legal fields, legal reasoning involves analysing the case through a real world perspective also. Along with the statutory rules, various heuristics imposed by culture, region, conventions and the experience of judges are also to be considered while making the decision. Given the case proceedings/current fact situation a highly structured legal reasoning system to analyse the case and thereby predict the most probable judgement based on the statute and discretion of the judge is proposed in this paper. It is hoped that the proposed legal counselling system will be of use to our society in the following ways.

- (1) The system, by its ability to predict in advance the most probable outcome in a given case, will enable individual clients to decide about the advisability or otherwise of entering into a legal dispute in a given situation. This in turn will lead to reduced workload on the considerably over-burdened courts.

- ## 2. Proposed legal system

(1) Technical information consists of particulars of sections of the relevant act invoked in dealing with the case, i.e. the ingredients and evidence level at which each of the ingredients has been established. This information regarding a specific case can be represented as an instance of the section's decision lattice (D-lattice).



- (2) Nontechnical information or the real world information of the case, such as the details of how and why the crime was committed can be represented as instances of the corresponding commonsense lattices (C-lattice).
- (3) Formal general information regarding the sentential details of each section is represented as a sentencing lattice (S-lattice) and it is of static nature.

When the user interacts with the system, the *shell* collects the case details through a question-answering session. The shell used the C-lattice instances to accomodate the details of the real world information of the present case. *Evidence estimator & D-lattice filler* gets technical information of the present case from the shell, and prepares the D-lattice instance representing the case in view of the relevant section. *Case strength evaluator* evaluates the corresponding D-lattice instance to measure the strength of a given case in accordance with the statute. The *discretion module* accomodates the experience-based real world knowledge of legal professionals as nontechnical heuristics. *Credibility evaluator* applies these heuristics on the C-lattice instances of the case to determine the credibility of the case. *Decision maker* suggests a decision on whether the accused has to be convicted or not based on the combined effect of strength and credibility of the case.

The judgement of a case includes the decision whether to convict or not as well as the sentence to be undergone by the accused if necessary. If a decision to convict the accused is taken, the decision-maker enables the sentencing module. *Severity evaluator* processes the C-lattice instances of the present case to get a severity measure of the crime committed. Based on this measure, punishment will be meted out to the accused in accordance with the sentential norms contained in the relevant S-lattice. According to the norms provided by the S-lattice and the severity of the present case, sentencing will be made by the *sentencing module*.

Since human reasoning is being simulated in a specific domain, the system becomes an expert system (Keller 1987) as its decision-prediction performance tends to that of a human expert. In any case, this system has been developed in an attempt to provide intelligent professional assistance to legal professionals and offers intelligent support to busy legal professionals while applying the regular domain specific techniques in case analysis so that they can concentrate better on critical aspects of cases. In this paper the processing of nontechnical knowledge to estimate the credibility of a case is dealt with in detail.

### 3. Knowledge structuring

Nontechnical knowledge of a case involves information regarding the details of the crime. This knowledge should be organised as a hierarchical system so that the details of higher level objects can be elaborated at lower levels. A highly accepted knowledge structure that can represent a complex object as a hierarchical system is FRAME (Rich & Knight 1991).

#### 3.1 Frames

Frames are one of the highly accepted knowledge representational formalisms in the field of AI, in particular in computer vision and natural language understanding. A frame represents a complex stereotypical object/occurrence and its slots represent the stereotypical aspects of the object. A slot can contain another frame or an atom as

its value at any of its various associated facets. The facets act as directives to the inferencing mechanism. An instance of a frame represents a specific object/occurrence and each of its slots can accommodate the particulars of the associated aspect of the specific object. In case of the absence of an aspect in an instance frame, it can inherit that aspect from its class frame. In case of the absence of an aspect in a class frame, it can inherit that aspect from its nearest ancestor. This value inheritance (Tichy 1987) property allows frames to avoid redundancy and to be concise. The value inheritance property makes the frames suitable for natural language understanding etc., where implicit knowledge retrieval is essential. The proposed legal system does not need the value inheritance since all individual facts of the case should be established explicitly. At the stage of predicting/making judgement the legal domain is a closed world and no attempts to establish the missing facts are allowed. Hence, the procedural attachment feature of frames in terms of demons etc. is also not necessary. Rather, the hierarchical knowledge structuring aspects of the frame suggest a new knowledge structure called LATTICE to represent the informal knowledge of legal domain.

### 3.2 LATTICE

A class of objects/occurrences with a predefined set of attributes can be represented as a lattice. The specific information regarding a particular object/occurrence can be represented as an instance of the class lattice. The values of an attribute of the instance lattice can be filled, if and only if the corresponding class lattice supports that attribute (i.e. if it is a relevant attribute). Instead of unidimensional attributes, the lattice has two-dimensional attributes for the following benefits.

- (1) Two-dimensional attributes make the lattice more expressive and nearer to the natural way of representing legal information.
- (2) Due to the modularity derived by the two-dimensional attribute lattice, it is preferred by domain/legal experts. Hence, knowledge acquisition is convenient.
- (3) a. Conversion of the domain expert's knowledge into internal knowledge structures is simpler for the knowledge engineer.  
b. Checks for completeness and making modifications to the existing knowledge are more convenient due to the modularity.

The value of an attribute of an instance lattice can either be an atomic value or an instance of another lattice as dictated by the nature of the attribute.

**3.3a Knowledge representation:** Nontechnical information of a case involves details of the case in layman's view. This knowledge can be represented using various C-lattices. The set of C-lattices to represent theft cases are as follows.

- (a) *Case-Ref*: This lattice is at the topmost level in the lattice system. This has to be accessed by the reference number of the case.
- (b) *Accused-name*: This lattice gives the details of the accused in this case. All relevant known information of the accused should be filled into various attributes of this lattice.
- (c) *Execution-Ref*: This lattice accommodates the details of the commitment of the crime. These details are in turn structured into the three lattices – event-no, abettors-name, item-name.
- (d) *Event-no*: This lattice represents the details of a particular event such as when and where the event happened.

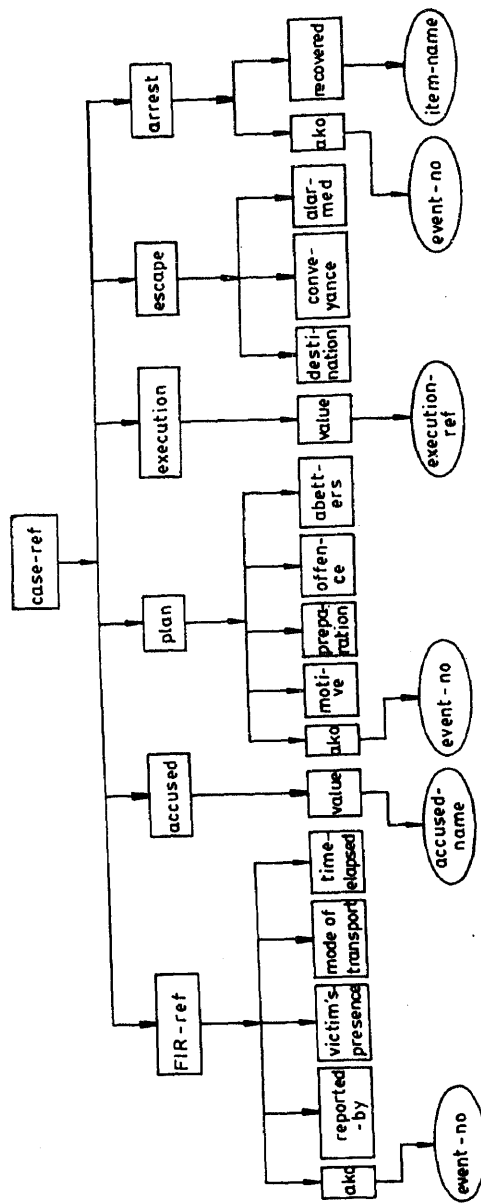


Figure 2a. Case-reference lattice.

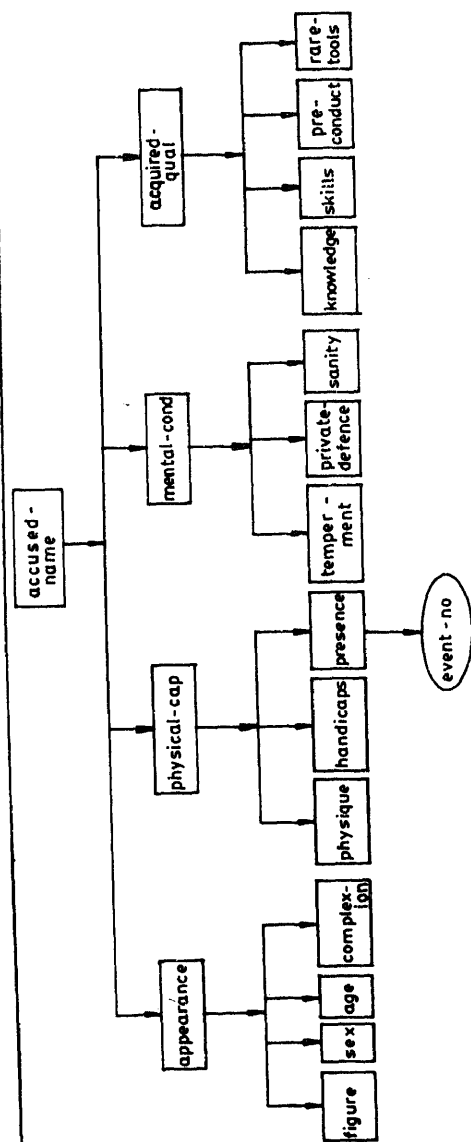


Figure 2b. Accused-name lattice.

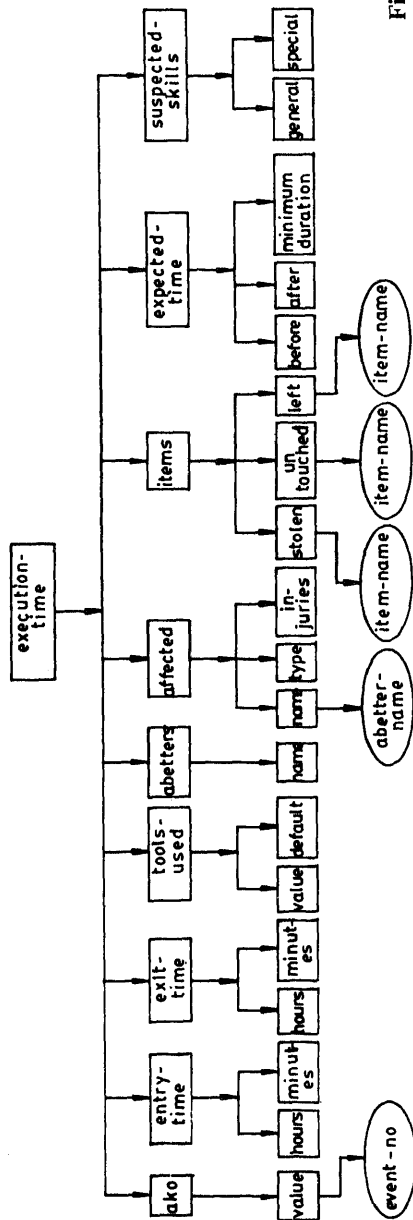


Figure 2c. Execution-time lattice.

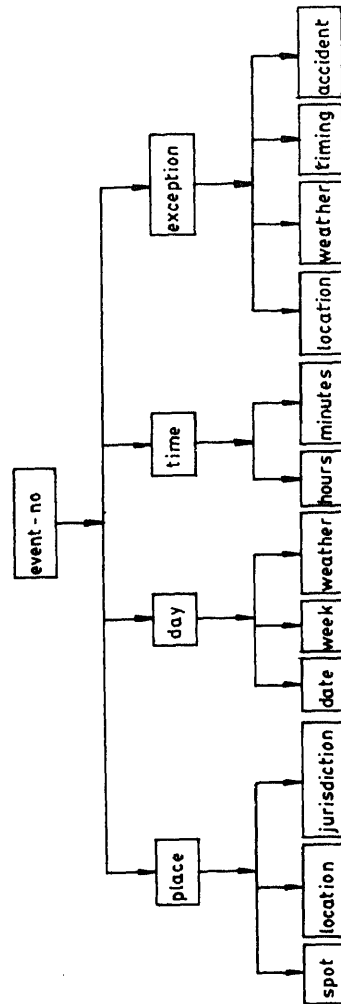


Figure 2d. Event-number lattice.



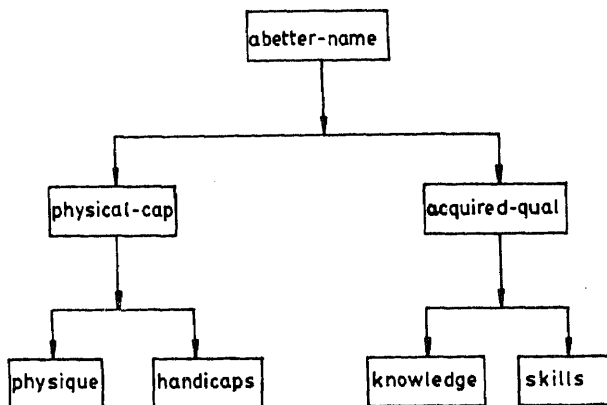


Figure 2e. Abetter-name lattice.

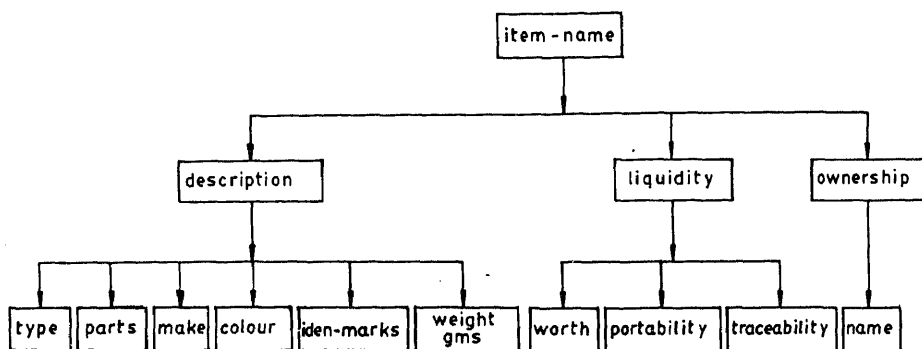


Figure 2f. Item-name lattice.

(e) *Abetter-name*: This represents the relevant capabilities of the abettors of the case.

(f) *Item-name*: It represents the characteristics of a particular item of interest.

These C-lattices are depicted as shown in figure 2.

**3.3b C-lattice operators:** C-lattices provide the structure for organising the real world/nontechnical knowledge of a particular case. Each of these provides a general structure for a chunk of relevant nontechnical knowledge. Several functions were developed in Common-LISP to operate with these lattices. The operations needed to store and retrieve the details of a case are as follows:

- (1) (Intro-instance <ref-no> case-ref): This function generates an instance of case-ref lattice and identifies it with <ref-no>.
- (2) (Ct-put <lattice-id> <attribute-path> <value>): This function is called while storing the details of a case. The value of the detail is stored in the identified lattice at the location according to the <attribute-path>. While storing, the function checks the relevancy of the attribute-path. Automatic introduction of the value as an instance of its compatible lattice is done through this function.
- (3) (Ct-remove <lattice-id> <attribute-path> <value>): This function can be used if a particular value of an attribute is found to be wrong and has to be deleted. The value will be deleted from the list of values of the attribute of the identified lattice.

- (4) (Ct-update <lattice-id> <attribute-path> <value>): This function can be used to overwrite the previous value of an attribute with a new value. When this function is called the <value> will be stored as a single value of <attribute-path> of the <lattice-id>.
- (5) (Ct-get <lattice-id> <attribute-path>): This function will be used to fetch/retrieve the list of values of <attribute-path> of lattice identified.
- (6) (Ct-removelatt <lattice-id>): This function can be used to delete lattices that were introduced as sub-structures to the lattice-id in a cascaded way. This function will be of use in cases of withdrawal of a case or cases that are finalised.

### 3.4 Discretion module

C-lattice instances associated with a case can be processed with the discretion module to evaluate the credibility of the case. The discretion module consists of heuristic knowledge of judges. This heuristic knowledge is represented procedurally over the C-lattice operators. Various chunks of heuristic knowledge are represented as individual 'rules' and a rule either supports or opposes the guilt of the accused. Some of the heuristics useful for dealing with theft cases have been implemented in our legal system. They are as follows.

#### RULE 1

IF	the belongings of the accused are found at the scene of occurrence of the crime
UNLESS	all of them are explained reasonably
CONCLUDE	to increase the credibility of the charge/commission of the offence of theft.

#### RULE 2

IF	the accused takes away less valuable items apparently leaving high valued items
UNLESS	there is a threat of being captured on the spot
	OR
	the portability of the stolen item is more than that of items untouched
	OR
	the untouched items are easily traceable
CONCLUDE	to reduce the credibility of the case.

#### RULE 3

IF	the accused who is old/child/female forced the stronger victims
UNLESS	the accused is supported by a strong weapon or a chemical or an abetter
CONCLUDE	to reduce the credibility of the case.

#### RULE 4

IF	the presence of the accused is recorded at a place other than the scene of occurrence at reasonably the same time that the crime was committed (alibi)
----	--

UNLESS journey by any viable fast transport makes it possible to reach the destination within the stipulated time  
 AND  
 the accused is healthy and capable of doing such a journey  
 CONCLUDE to make the credibility of the case zero.

**RULE 5**

IF the accused is not sound physically/mentally at the time of commission of the crime  
 UNLESS the experts certify his capability to perform all the required skills to commit the crime  
 OR  
 abettors can help him with those skills  
 CONCLUDE to reduce the credibility of the case to a greater extent.

**RULE 6**

IF time elapsed between entry and exit of the accused into the crime scene is less than the minimum expected duration of crime  
 UNLESS with the support of a familiar abettor or the accused himself is familiar with the scene of crime  
 CONCLUDE to make the credibility of the case zero.

**RULE 7**

IF the accused acquired/prepared a rare tool or vehicle that was used/suspected to be used while executing the crime  
 UNLESS he lost it well before the occurrence of crime  
 CONCLUDE to increase the credibility of the case to a greater extent.

**RULE 8**

IF the accused did not acquire the required special skills  
 UNLESS the skilled abettors helped him  
 OR  
 an effective preparation to take care of the situation is recorded  
 CONCLUDE to reduce the credibility of the case.

**RULE 9**

While comparing the recovered items with the stolen items

(a) IF some recovered items were found identical in all aspects to the stolen items  
 UNLESS the accused proves his right of possession/ownership on all those items  
 CONCLUDE to increase the credibility of the case  
 (b) IF all recovered items differed from the stolen items in one way or the other  
 CONCLUDE to reduce the credibility of the case.

#### 4. Credibility evaluator

Credibility is a positive real number associated with each case to represent the 'believability' of the case. For the sake of unbiased evaluation, the credibility of the case should be initialised to unity which neither supports nor opposes the guilt of the accused prior to evaluation. Then credibility evaluator selects the applicable discretion rules and executes them in an order dictated by the offence involved. In this process, the credibility of a case may increase/decrease in accordance with the execution of rules that support/oppose the guilt of the accused. The resultant credibility will be returned as a real number. If the resultant credibility is more than unity the accused is more likely to be convicted and if it is less than unity he may be acquitted. Credibility suggests the judgement in view of nontechnical information of the case. A sample session with credibility evaluator is given in appendix B.

#### 5. Conclusions

Computer-based legal systems have to progress a long way to aid legal reasoning rather than legal information retrieval. The existing legal consultation systems are aimed at certain specific civil cases and a few of these systems attempt criminal cases. The distinctive features of criminal cases as against civil cases is the increased effectiveness of nontechnical matters in reaching the judgement. In this paper a model of a judgement prediction system has been proposed. This model aims at analysing a specific criminal case through technical as well as nontechnical perspectives and accordingly suggest the judgement. Co-accused cases are not considered in the present model. The components of the model to analyse the case through nontechnical perspectives are implemented in Common-LISP on the APOLLO, NEXUS 3500 under aegis. Though the subsystem developed is limited to handling theft cases, it can be extended to most other criminal cases.

The authors would like to thank Prof P V Ratnam, of the Department of Electrical Engineering for providing the APOLLO, NEXUS 3500 system. The first two authors benefited from stimulating discussions with Mr. C N V D Sastry, an advocate.

#### Appendix A. Glossary of terms.

Accused:	a person against whom a case is filed in a criminal court;
Acquittal:	decision by the court after finding the accused not guilty;
Conviction:	decision by the court after finding the accused guilty;
Defendant:	one who defends himself in a court of law;
Ingredient:	an essential requirement;
Plaintiff:	one who initiates (files) a case in a civil court;
Precedent:	a binding case;
Section:	a provision in a statute;
Sentencing:	order imposing punishment by the judge;
Statute:	an act passed by the legislature duly assented.

## Appendix B. Sample sessions.

Nontechnical information processing to estimate the credibility of theft cases is illustrated through the following sample session with the system.

### Case-1

#### Description of the case-1

On 29th June, 1992, Monday, at around 2.30 a.m. a theft occurred in the house of Sri Ramesh, situated at Lawsons Bay, Visakhapatnam. While the inmates were sleeping the accused entered the house through a ventilator with a rope, while an abetter waited outside the house. The accused threatened the inmates with a sharp knife and stole a gold chain worth Rs. 10,000/- weighing 30 g, a gold ring worth Rs. 3,000/- weighing 10 g bearing the identification mark 'Th' on it, and cash equal to Rs. 5,000/-. When the watchman (*gurkha*) approached the house, the abetter heard him, signalled to the accused through a window and both of them escaped. Four silver plates worth Rs. 16,000/- weighing 2000 gms. were left untouched. Two days later, Pal and Raheem were arrested in Kakinada while they were trying to sell a gold chain (weighing 29 g) and a ring (weighing 10 g) which were similar to the stolen articles. The victims of the offence recognised Pal as the offender. It was found that the rope left at the scene of the crime was bought by Pal two days prior to the day of the crime. The accused Pal (30) is a strong man. Though he is dumb and deaf, he is skilled in climbing heights with a rope. The abetter Raheem is skilled in liquidating gold articles.

#### C-LATTICES REPRESENTING CASE-1 :

```
(C1S-380 (IS-A (VALUE (CASE-REF)))
```

```
  (ACCUSED (VALUE (PAL)))
```

```
  (EXECUTION (VALUE (EX-1)))
```

```
  (ARREST (RECOVERED (RING2) (CHAIN2))
```

```
    (AKO (EV-2)))
```

```
  (ESCAPE (ALARMED (GHORKA))))
```

```
(PAL (IS-A (VALUE (ACCUSED-NAME)))
```

```
  (APPEARANCE (AGE (30))
```

```
    (SEX (MALE)))
```

```
  (PHYSICAL-CAP (PHYSIQUE (STRONG))
```

```
    (HANDICAPS (DUMB-AND-DEAF)))
```

```
  (ACQUIRED-QUAL (SKILLS (CLIMBING-WITH-ROPE))))
```

```
(EX-1 (IS-A (VALUE (EXECUTION-REF)))
```

```
  (AKO (VALUE (EV-1)))
```

```
  (ABETTERS (NAME.(RAHEEM)))
```

```
  (TOOLS-USED (VALUE (ROPE1) (KNIFE)))
```

(SUSPECTED-SKILLS (GENERAL (RUNNING))  
                                  (SPECIAL (CLIMBING-WITH-ROPE)  
  (LIQUIDATING-GOLD)))  
(AFFECTED (TYPE (MALE) (FEMALE)))  
(ITEMS (STOLEN (CHAIN1)  
                                  (RING1)  
                                  (CASH1))  
          (UNTOUCHED (SILVER-PLATES))  
          (LEFT (ROPE1))))

(EV-1 (IS-A (VALUE (EVENT-NO)))  
      (PLACE (SPOT (DWELLING-HOUSE))  
          (LOCATION (LAWSONSBAY))  
          (JURISDICTION (VISAKHAPATNAM)))  
(DAY (DATE (29-6-92))  
      (WEEK (MONDAY)))  
(TIME (HOURS (2))  
      (MINUTES (30))))

(CHAIN1 (IS-A (VALUE (ITEM-NAME)))  
          (DESCRIPTION (TYPE (ARMAMENT))  
                          (MAKE (GOLD-90))  
                          (WEIGHT-GMS ((30 0.95))))  
(LIQUIDITY (WORTH (10000))  
          (PORTABILITY (VERY-HIGH))))

(RING1 (IS-A (VALUE (ITEM-NAME)))  
          (DESCRIPTION (TYPE (ARMAMENT))  
                          (MAKE (GOLD-90))  
                          (WEIGHT-GMS ((10))  
                          (IDEN-MARKS (TH)))  
(LIQUIDITY (WORTH (3000))  
          (PORTABILITY (VERY-HIGH))))

(CASH1 (IS-A (VALUE (ITEM-NAME)))  
          (LIQUIDITY (WORTH (5000))  
          (PORTABILITY (HIGH)))  
(DESCRIPTION (TYPE (MONEY))))

(SILVER-PLATES (IS-A (VALUE (ITEM-NAME)))  
          (LIQUIDITY (PORTABILITY (MEDIUM))  
          (WORTH (16000))  
(DESCRIPTION (WEIGHT-GMS (2000))))

(ROPE1 (IS-A (VALUE (ITEM-NAME)))  
(OWNERSHIP (NAME (PAL))))

(RAHEEM (IS-A (VALUE (ABETTER-NAME)))  
(ACQUIRED-QUAL (SKILLS (LIQUIDATING-GOLD))))

(EV-2 (IS-A (VALUE (EVENT-NO)))  
(DAY (DATE (01-7-92)))  
(PLACE (JURISDICTION (KAKINADA))))

(CHAIN2 (IS-A (VALUE (ITEM-NAME)))  
(DESCRIPTION (TYPE (ARNAMENT))  
(MAKE (GOLD-90))  
(WEIGHT-GMS (29))))

(RING2 (IS-A (VALUE (ITEM-NAME)))  
(DESCRIPTION (TYPE (ARNAMENT))  
(MAKE (GOLD-90))  
(WEIGHT-GMS (10))  
(IDEN-MARKS (TH))))

CASE 1. Evaluation follows in context 1.

>(evaluate`C1S-380)

ROPE 1 belonging to accused was found at the scene of occurrence.

Is this reasonably explained?

Indicate y/n. n

Does the deformity (DUMB-AND-DEAF) allow the accused to perform EACH and EVERY ONE of the following tasks (even with the help of RAHEEM)?

(RUNNING, CLIMBING-WITH-ROPE)

Consult the experts and accordingly indicate y/n. y

It is assumed that the weight of RING 1 is exact.

Did the accused prove his ownership/right of possession regarding each of the following items?

(CHAIN 2, RING 2)

Please indicate y/n. n

1:5625 is the value of credibility for the present case C1S-380.

THANK-YOU!

CASE 1. Evaluation follows in context 2.

>(evaluate`C1S-380)

ROPE 1 belonged to accused was found at the scene of occurrence.

Is this reasonably explained?

Indicate y/n. y

Does the deformity (DUMB-AND-DEAF) allow the accused to perform EACH and EVERY ONE of the following tasks (even with the help of RAHEEM)?  
(RUNNING, CLIMBING-WITH-ROPE)

Consult the experts and accordingly indicate y/n. y

It is assumed that the weight of RING 1 is exact.

Did the accused prove his ownership/right of possession regarding each of the following items?

(CHAIN 2, RING 2)

Please indicate y/n. y

1 is the value of credibility for the present case C1S-380.

THANK-YOU!

## CASE 2

### Description of case 2.

On 2nd August, 1992, Sunday, at 8-15 p.m. a theft occurred in the house of Reddy, situated at Banjara Hills, Hyderabad. Reddy returned from his office with a briefcase containing one lakh rupees in his blue Maruti-92 car. After he relaxed for 5 minutes he found that a man of 25 years was driving away in his car and immediately noticed that the briefcase containing the cash was missing. Through investigation it was found that Geetha, the maid servant in the house, had dropped the briefcase and the car keys to help the accused. Three days later, one Rao was arrested with a similar red Maruti car in Warangal. The accused produced an alibi showing evidence that he was consulting a doctor in Tata Hospital, Bombay, on the day of the theft at 5-30 p.m.

### C-LATTICES REPRESENTING CASE-2 :

(C2S-380 (IS-A (VALUE (CASE-REF))))

(ACCUSED (VALUE (RAO)))

(EXECUTION (VALUE (EX-2)))

(ARREST (AKO (EV-22)))

(RECOVERED (CAR21)))

(EX-2 (IS-A (VALUE (EXECUTION-REF))))

(ENTRY-TIME (HOURS (8)))

(MINUTES (13)))

(EXIT-TIME (MINUTES (15)))

(HOURS (8)))

(AKO (VALUE (EV-20)))

(ITEMS (STOLEN (CASH20) (CAR20)))

(AFFECTED (NAME (REDDY)))

(TYPE (MALE)))

(ABETTERS (NAME (GEETHA)))

(SUSPECTED-SKILLS (GENERAL (VISION)))

(SPECIAL (CAR-DRIVING)))

(EXPECTED-TIME (MIN-DURATION (5)))



(EV-20 (IS-A (VALUE (EVENT-NO)))  
 (PLACE (SPOT (HOUSE))  
 (LOCATION (BANJARA-HILLS))  
 (JURISDICTION (HYDERABAD)))  
 (DAY (DATE (2-8-92))  
 (WEEK (SUNDAY)))  
 (TIME (HOURS (8))  
 (MINUTES (15))))

(CASH20 (IS-A (VALUE (ITEM-NAME)))  
 (LIQUIDITY (WORTH (100000))  
 (TRACEABILITY (LOW))))

(CAR20 (IS-A (VALUE (ITEM-NAME)))  
 (DESCRIPTION (TYPE (VEHICLE))  
 (MAKE (MARUTI-92))  
 (IDEN-MARKS (701284))  
 (COLOUR (BLUE)))  
 (LIQUIDITY (WORTH (120000))  
 (TRACEABILITY (HIGH)))  
 (OWNERSHIP (NAME (REDDY))))

(GEETHA (IS-A (VALUE (ABETTER-NAME)))  
 (ACQUIRED-QUAL (KNOWLEDGE (INMATE))))

(RAO (IS-A (VALUE (ACCUSED-NAME)))  
 (APPEARANCE (AGE (25))  
 (SEX (MALE)))  
 (PHYSICAL-CAP (PRESENCE (EV-21)))  
 (ACQUIRED-QUAL (SKILLS (CAR-DRIVING))))

(EV-21 (IS-A (VALUE (EVENT-NO)))  
 (PLACE (SPOT (TATA-MEMORIAL-HOSPITAL))  
 (LOCATION (DADAR))  
 (JURISDICTION (BOMBAY)))  
 (DAY (DATE (2-8-92))  
 (TIME (HOURS (5))  
 (MINUTES (30))))

(CAR21 (IS-A (VALUE (ITEM-NAME)))  
 (DESCRIPTION (TYPE (VEHICLE))  
 (MAKE (MARUTI-92))  
 (IDEN-MARKS (701284))  
 (COLOUR (RED))))

(EV-22 (IS-A (VALUE (EVENT-NO)))  
 (PLACE (JURISDICTION (WARANGAL)))

CASE 2. Evaluation follows in context 3.

(>evaluate' C2S-380)

What is the distance in kilometres between HYDERABAD and BOMBAY?

750

Can the accused fly between HYDERABAD and BOMBAY?

Indicate y/n. y

Check whether a flight took off at BOMBAY on 2-8-92 after 6'O clock and reached HYDERABAD before 8.

Please indicate y/n. n

C2S-381 INVALID

The court believes the alibi is reasonable.

0 is the value of credibility for the present case C2S-380.

THANK-YOU!

CASE 2. Evaluation follows in context 4.

(>evaluate' C2S-380)

What is the distance in kilometres between HYDERABAD and BOMBAY?

750

Can the accused fly between HYDERABAD and BOMBAY?

Indicate y/n. y

Check whether a flight took off at BOMBAY on 2-8-92 after 6'O clock and reached HYDERABAD before 8.

Please indicate y/n. y

Is there a possibility to change the colour of CAR 21?

Indicate y/n. y

Did the accused prove his ownership/right of possession regarding each of the following items?

(CAR 21)

Please indicate y/n. n

1:25 is the value of credibility for the present case C2S-380.

THANK-YOU!

## References

- Ashley K D 1991 Reasoning with cases and hypotheticals in HYPO. *Int. J. Man Machine Studies* 34: 753-796
- Ashley K D, Rissland E L 1988 Dynamic assessment of relevancy in a case-based reasoner. *Proceedings of the Fourth Conference on Artificial Intelligence Applications*, California. pp. 208-214

- Gardner A L 1987 *An artificial intelligence approach to legal reasoning* (ed.) Bradford Book (Cambridge, MA: The MIT Press)
- Hafner C D 1987 Conceptual organisation of case law knowledge bases. *Proceedings of the First International Conference on AI and Law* (New York: ACM)
- Keller R 1987 *Expert system technology: Development and application* (Englewood Cliffs, NJ: Prentice-Hall)
- McCarthy T 1980 The TAXMAN Project: Towards a cognitive theory of legal argument. *Computer science and law: An advanced course* (ed.) B Niblett (New York: Cambridge University Press)
- Mehl L 1958 Automation in the legal world. *Proceedings of Symposium on Mechanisation of Thought Processes* (Teddington, London: Natl. Phys. Lab.)
- Rich E, Knight K 1991 *Artificial intelligence* 2nd edn (New Delhi: Tata McGraw-Hill)
- Srivastava S K 1991 Case-based systems in law: A survey. Project report, Department of Electronics, New Delhi
- Tichy W F 1987 *IEEE Comput.* 20(11): 43-54

## A prototype expert system for interpretation of remote sensing image data

L CHANDRA SEKHARA SARMA<sup>1</sup> and V V S SARMA<sup>2</sup>

<sup>1</sup> ISRO Computer Office, Indian Space Research Organisation HQ, Bangalore 560 094, India

<sup>2</sup> Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India

E-mail: <sup>1</sup>lcs@isro.ernet.in; <sup>2</sup>vvs@csa.iisc.ernet.in

**Abstract.** Automated image interpretation systems of remotely sensed images are of great help in the present scenario of growing applications. In this paper, we have critically studied visual interpretation processes for urban land cover and land use information. It is observed that the core activity of interpretation can be described as plausible combinations of pieces of evidential information from various sources such as images, collateral data, experiential knowledge and pragmatics. Interpretation keys for the interpretation of standard false colour composites are considered to be tone/colour, pattern, texture, size, shape, association, relief and season. These interpretation keys encompass the spectral, spatial and temporal knowledge required for image interpretation. Our focus is on a knowledge-based approach for interpretation of standard false colour composites (FCC). Basic information required for a knowledge-based approach is of four types viz., spectral, spatial, temporal and heuristic. Generic classes and subclasses of image objects are identified for the land use/land cover theme. Logical image objects are conceptualised as region/area, line and point objects. An object-oriented approach for the representation of spectral and spatial knowledge has been adopted. Heuristic information is stored in rules. The Dempster–Shafer theory of evidence is used to combine evidence from various interpretation keys for identification of generic class and subclass of a logical image object. Analysis of some Indian Remote Sensing Satellite images has been done using various basic probability assignments in combination with learning. Explanation facility is provided by tracing the rules fired in the sequence.

**Keywords.** Remote sensing; image interpretation; false colour composite; interpretation keys; expert systems; domain objects; Dempster–Shafer theory; uncertainty handling.

### 1. Introduction

The need for automated image interpretation systems with expert-level performance has been long felt. Although the intent of computer-assisted digital image classification

is to generate thematic maps using more quantitative methods, visual interpretation is still indispensable for attaining expert-level performance. To a large extent, this is due to the inadequacies of digital classification such as lack of temporal, spatial and neighbourhood knowledge. The need for expert-level performance in image interpretation has brought in a paradigm shift from domain independent statistical methods to domain specific knowledge-based techniques (Argialas 1990). The core activity of interpretation can be described as a plausible combination of pieces of evidential information from various sources such as characteristic features of image objects, domain-specific knowledge, collateral data and pragmatics. This activity is more in the nature of explorative and qualitative reasoning in the line of artificial intelligence (AI) and expert systems (ES). AI and ES techniques have contributed powerful and flexible methodologies to represent domain-specific knowledge and heuristic problem-solving knowledge in the domain of image interpretation which are often declarative in nature. While many knowledge-based methodologies combining AI, pattern recognition and image analysis have been proposed by quite a few researchers in the recent past (Wang & Newkirk 1988; Schowengerdt & Wang 1989), there are hardly any systems which consider knowledge from spectral, spatial and temporal domains together for interpreting an image.

We have developed a prototype expert system for the interpretation of false colour composites (FCC) of IRS-1A (Indian Remote-Sensing Satellite) for land use/land cover categorization theme, using GC LISP on a personal computer. Various logical components of this system are given in figure 1. The visual interpretation key developed by the National Natural Resources Management Systems (NNRMS) Office, Department of Space, has been used as a basis for our knowledge-based approach, which covers the required knowledge from all the three domains mentioned before. Image interpretation activity is viewed as a data fusion activity in which the sources of evidence are features such as colour, texture, pattern, size, shape, association, relief and season. Knowledge is represented in property lists of GC LISP in the form of objects and rules. Knowledge organization is hierarchical (two-level) and control sequence is sequential. Reasoning for identification of logical image objects is done using Dempster's

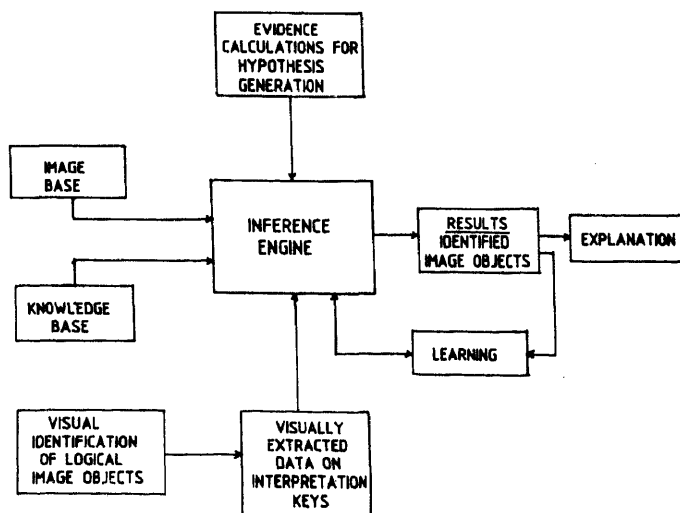


Figure 1. Various logical components of the expert system.

combination rule for combining evidential information from the various features mentioned above.

## **2. Formulation of knowledge base**

### *2.1 Knowledge elicitation*

For knowledge elicitation, we have conducted focused and structured interviews with experts available at the NNRMS office based on the interpretation key developed by a team of experts in the theme of land use/land cover categorization. Part of the interpretation key is shown in table 1. Imprecision and vagueness in the description of feature values is recognized and the experts agree with it as it is because of inherent fuzziness in human expression. Knowledge was elicited about variation in weightages to be allocated to feature values based on domain, rules for interpretation and conflict resolution, in the case of different results being obtained with the same subset of feature values. The term 'interpretation key' is used in the sense of 'feature' in further discussions.

### *2.2 Types of the knowledge*

We identify the types of knowledge and data in the domain of image interpretation and the relevant goals, as shown in table 2. This analysis provides support in designing a knowledge-based system, in deciding the choice of the knowledge-representation and uncertainty-handling schemes and in performance analysis (Hayes-Roth 1989).

### *2.3 Uncertainty handling*

Image interpretation involves decreasing the local ambiguity and merging the pieces of knowledge (associated with the interpretation keys) into a unique interpretation. The disambiguation process calls for handling uncertainty in the domain of image interpretation. We choose to accept the confidence factors provided by the user, which are representatives of the user belief in expressing the values of the corresponding features of an image object. Reasoning is done to identify image objects based on the feature values and associated certainty factors.

### *2.4 Steps involved in the system design*

To model and implement a knowledge-based system for interpretation of satellite imagery, four levels are identified. They are the conceptual, representation, reasoning and idealization levels as shown in figure 2.

In our system spectral, spatial and temporal knowledge embedded in the interpretation keys together generate a hypothesis based on image-domain, scene-domain specifications (shown in figure 3) and on the user's confidence in the description of the feature values. This hypothesis may suggest a subset of identification names, which is further refined using real-world knowledge and heuristics to label an image object.

**Table 1.** Land use/land cover interpretation key using satellite remote sensing imagery.

Sl. No.	Land use/land cover category	Tone/colour	Size	Shape	Texture	Pattern
01	Built-up land	Dark bluish green in the core and bluish on the periphery	Small to big	Irregular & discontinuous	Coarse & mottled	Clustered to scattered & non-contiguous
02	Transportation	Very dark to dark bluish green, light yellow for minor roads, red if vegetation along the road	Small in width for roads and narrow for rail	Regular with straight/sharp and smooth curves	Smooth to fine	Linear to sinuous & contiguous
03	Crop land	Bright red to red	Varying in size	Regular to irregular	Medium to smooth	Contiguous to non-contiguous
04	Fallow land	Yellow to greenish blue (depending on soil type and moisture)	Small to large	Regular to irregular	Medium to smooth	Contiguous to non-contiguous
05	Plantation (agriculture)	Dark red to red	Small to medium	Regular with sharp edges	Coarse to medium	Dispersed contiguous
06	Evergreen/Semi-ever-green forest	Bright red to dark red	Varying in size	Irregular, discontinuous	Smooth to medium depending on crown density	Contiguous to non-contiguous
07	Deciduous forest	Dark red to red	Varying in size	Irregular, discontinuous	Smooth to medium depending on crown density	Contiguous to non-contiguous

Location	Association	Season	Remarks
Plains, plateaus, on hill slopes, deserts, water-front, road, rail, canal etc.	Surrounded by agricultural lands, forest cover, wastelands, network of rivers, roads, and rail etc.	October to March	Built-up land can be of big or small size settlements, industrial structures, buildings or any other artifact, physical spread or sprawl along with density of transport network are useful surrogates to classify it as urban or rural. Perceptible land transformation can be noticed around built-up land
On all types of terrain, across water bodies, agricultural lands connecting settlements	Settlement nodes, amidst and around built-up developed areas etc.	October to March	Provides connectivity linkages between settlements and accelerates development. Road, rail and canal vary in dimension and importance. Can be mapped in detail using infrared bands and higher spatial resolution data. Forms part of non-agricultural use
Plains, hill slopes, valleys, cultivable wastelands etc.	Amidst irrigated (canal, tank, well etc.) and unirrigated (rainfed/dry farming) arable lands, proximity to rivers/streams etc.	June to September and October to April	Consists of different crops grown in different seasons under different farming and land-tenural systems. Mixed and multiple cropping patterns generate mixed spectral response on the images
Plains, valleys uplands etc.	Amidst crop land as harvested agricultural fields etc.	January to December	Consists of different arable lands left uncultivated as seasonal/temporary fallows for less than a year and as permanent fallows up to 5 years or more because of diverse reasons. Fallow land devoid of vegetation, accelerates erosion
Plains, foot hills and uplands	Dry lands or unirrigated lands, uplands occasionally amidst crop land, proximity to rivers and on gentle hill slopes	January to December	Agricultural plantations consist of a variety of trees, orchards and groves. These occur throughout the year and are seen very prominently on the imagery. Those occurring in the forest areas (but outside the notified forest areas) are also treated as plantations like coffee, tea, arecanut etc.
High relief mountain/hill tops and slopes and within notified areas	High relief/slopes exposed to very heavy rainfall zones.	January to December	These are closed (40% tree cover) or high density forest cover of conifers and other broad leaved forest trees. These coincide with the zones of high rainfall and relief. They provide shelter to wildlife and livestock. They influence the climate and water regime and protect the environment
Medium relief mountains/hill slopes and within notified areas	Different forest types/sub-types of species which shed leaves	January to April	These are broad-leaved tropical forests which seasonally shed their leaves annually. Dry forest trees are subject to wild forest fires particularly during summer/autumn. These occur on the lower elevations and slopes rather than in the evergreen/semi-evergreen forests.

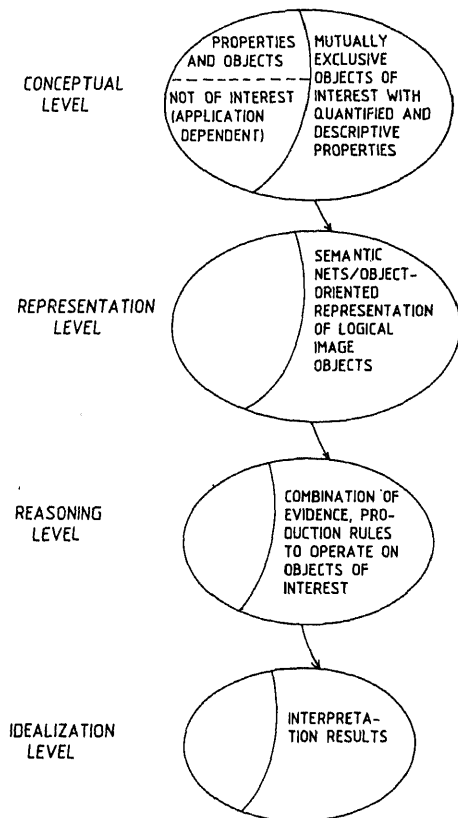


**Table 2.** Types of knowledge and data in image interpretation domain and relevant goals.

Type	Examples
<i>Objects</i>	
Domain objects	Road, rail, canal; plantation, tank/reservoir, settlements, industrial complexes, ships, tanks
Scene objects	Line objects, area objects, point objects
<i>Object attributes</i>	
Facts	Rivers do not cross each other
Defaults	Red colour indicates 'vegetation'
Factual rule	If the scene is urban-land and colour is white and shape is circular then the object is stadium
Heuristic	Assume black colour indicates a water body in the first instance
Fuzzy facts	Streams are with unstructured pattern and with 'somewhat narrow' starting and 'rather wide' ending
Fuzzy rule	If the texture of vegetation is 'smooth to medium' then it may be crop land
<i>Domain structures</i>	
Elementary structures	Point, line and area
Network structures	Drainage patterns
Group structures	Industrial sheds with housing colonies
<i>Prerequisites (data and data processing)</i>	
Spectral clarity	Enhancement, removal of noise in the pixels of image
Collateral material	Ground truth, toposheets, aerial photographs, geographic information system
Preference	FCC is preferred to B/W image for land cover categorization
<i>Problem-solving knowledge</i>	
Knowledge representation	Pixel oriented/vector representation/object oriented
Meta-knowledge	Examine line objects first for geological application; examine area objects first for land cover/land use applications
Heuristic meta-rule	Fire the rule with maximum confidence first
Combination of evidence	Additive, non-additive, <i>ad-hoc</i>
Uncertainty handling	Certainty factors, fuzzy calculus, belief measures
Conflict resolution	Interdependence of objects for recognizing 'association'
<i>Goals</i>	
Civilian	Monitoring man-land ratio estimates, water resource allocation etc.
Military	Troop movement observation, approachability and formation of regiments etc.

## 2.5 Knowledge representation

By and large, complex problems become tractable if one chooses the right level of abstraction, i.e. the set of appropriate terms in which to think about the domain. As an alternate approach to the image data base management systems which are found not suitable to handle feature-oriented image object knowledge, we conceptualize detectable image objects as point, line and region/area objects (Sarma & Sarma 1990) and adapt object-oriented approach for knowledge representation of image objects. Generic classes and corresponding subclasses in the domain of land use/land cover categorization are identified and two sample classes are shown in table 3.



**Figure 2.** Steps involved in knowledge-based interpretation system design.

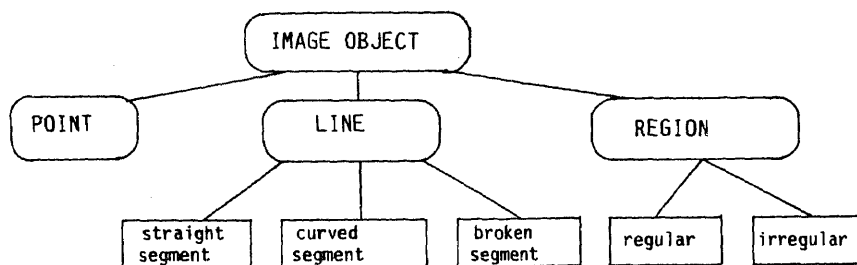
The class-subclass relation in a region object is represented using property lists in GC LISP as shown below.

```
(SETQ R2'((TYPE(VALUE REGION))
(NAME(CLASS AGRI-LAND)(SUB-CLASS CROP-LAND)
(COLOUR(VALUE((BRIGHTRED)(NORMLED)(LIGHT RED))))
(PATTERN(VALUE(CONTIGUOUS NON-CONTIGUOUS)))
(TEXTURE(VALUE(REGULAR IRREGULAR)))
(SIZE(VALUE(VARYING)))
(ASSOCIATION(SURROUNDED(VEGETATION
FOREST BUILT-UP-LAND) (CONTAINS (NIL) (SIDE-OF (RIVER
INFORMATION(VALUE
(DIFFERENTCROPSAREGROWNINDIFFERENTAREAS)))))
```

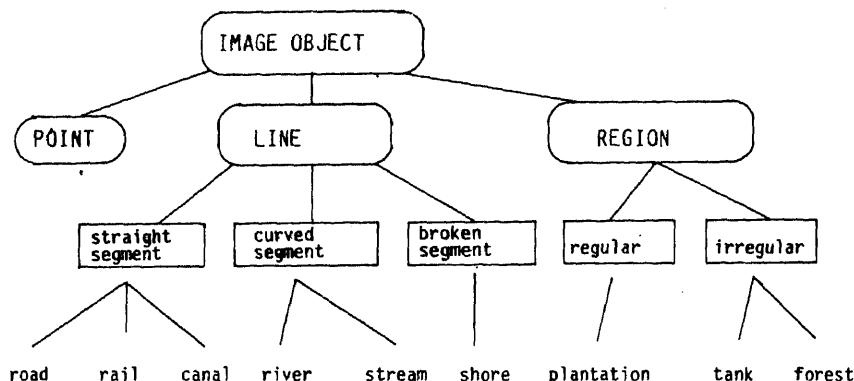
**Table 3.** Sample generic classes and corresponding subclasses in land use/land cover categorization.

Generic class	Subclass
Vegetation	Crop land Plantation Forestry
Built-up land	Settlements Urban/rural Industrial

(a)



(b)



(c)

- . Rivers do not cross each other
- . River tributaries flow into river or into shores
- . Exclusively river has a drainage basin
- . The inside area of a shore line is land iff its outside area is water; its inside is water iff its outside is land

Figure 3. Image domain (a) and scene domain (b) specifications, and real-world knowledge (c).

The knowledge about a line and a point object is represented interpretation key-wise in the form of rules as shown below.

Line object .

```

(SETQ LINE(APPEND LINE(RAIL ROAD RIVER CANAL)))
(SETQ RULE41'(COND (EQ COLOUR 'BLUISHGREEN))
(SETQ SETI'(ROAD))) ('T (SETQ SETI'NIL))))
  
```

Point object

```

(SETQ POINT (APPEND POINT'(INDUSTRIAL-SHED
BUILDING TREE SETTLEMENT)).
(SETQ RULE(APPEND RULE(QUOTE(RULE 2)
(SETQ RULE2'(COND ((EQ COLOUR 'RED)
(SETQ SETI'(TREE))) ('T(SETQ SETI'NIL))
  
```

An object-oriented approach for knowledge representation as shown above provides an environment for modular software design. Because of run-time binding facility in AI programming language LISP, it is possible to store and modify object data, facts and rules dynamically across different types and classes of logical objects of an image. Property-list structures are helpful in extending the existing knowledge base for future updates. Default knowledge and assumptions are stored in the form of rules, which take care of side-effects of the reasoning mechanism. Thus knowledge representation requirements are met effectively with an object-oriented approach.

It is observed that the time taken for the identification process for region objects based on property lists as facts in the knowledge-base is significantly high. Hence it is decided to go for rules structure in the knowledge base for line/point objects though the overall computation time complexity is the same (Sarma 1991).

### 3. Reasoning for identification

Satellite image interpretation activity involves analysis of an image to decrease the local ambiguity by fusing the pieces of knowledge associated with the interpretation keys into a unique interpretation. For example, each interpretation key may suggest one or more land cover categories; crop land is identified by the colour signature of bright red to red whereas plantation (agriculture) is identified by the colour signature dark red to red. This indicates that there is an overlap in the description of signature and hence colour alone may indicate two categories. Hence, we consider the other interpretation keys such as texture and pattern and fuse the knowledge from them with colour signature and bring out consensus to identify crop land and plantation. Thus image interpretation problem can be seen as a data fusion activity, in the sense that individual elements of an image object have to be associated in order to produce a comprehensive and unique interpretation. This approach helps in removing brittleness in decision.

#### 3.1 Basic theory

We have applied the Dempster-Shafer (D-S) theory of evidence to remote sensing satellite image interpretation for combining of evidence associated with the interpretation keys for the identification of a target object on a given false colour composite. In this theory, the belief in a proposition  $A$ , is expressed by a subinterval  $[s(A), p(A)]$  of the unit interval  $[0, 1]$ . The lower value  $s(A)$  represents the 'support' for that proposition and sets a minimum value for its likelihood. The upper value,  $p(A)$ , denotes the 'plausibility' of that proposition and establishes a maximum likelihood. 'Support' may be interpreted as the total positive effect a body of evidence has in a proposition, while 'plausibility' represents the total extent to which a body of evidence fails to refute a proposition. The degree of uncertainty about actual probability value for a proposition corresponds to the width of its interval.

#### 3.2 Formulation of representation of evidence

Let  $F$  be the mutually exclusive and exhaustive set of propositions in the domain, called frame of discernment or universe of discourse. Elements of the power set  $2^F$ , that is, subsets of  $F$  are the class of general propositions in the domain. Let  $N$  be

the number of features/interpretation keys based on whose values an object (point, line or area) is identified. These interpretation keys are considered as knowledge sources  $\{ks1, ks2, \dots, ksn\}$ .

### 3.3 Dempster's rule of combination

In this, each knowledge source distributes a unit belief across a set of propositions for which it has evidence. These propositions are referred to as focal elements of corresponding knowledge sources. The distribution is in proportion to the weight of that evidence as it bears on each proposition.

General formalism of the above description may be represented by a function

$$m_1 = \{A_i | A_i \subset F\} \rightarrow [0, 1],$$

where  $F$  is a mutually exclusive and exhaustive set of propositions in the domain.

Support for a proposition.

$$s_1(A) = \sum_{A_i \subset A} m_1(A_i).$$

$s_1(A)$  is also denoted by  $\text{Bel}(A)$  which indicates total belief of  $A$ .  $\text{Bel}(A)$  is called a belief function if it satisfies the following properties (Ng & Abramson 1990):

- (i) the belief in a null hypothesis is 0;
- (ii) the belief in  $F$  is 1;
- (iii) the sum of beliefs of  $A$  and  $\tilde{A}$  must be less than or equal to 1.

Total mass  $m(C)$ , combining masses from two sources  $ks1$  and  $ks2$  are combined using the formula (here  $C$  is a given subset of  $F$ )

$$m(C) = \begin{cases} 0 & \text{if } C \text{ is } \phi, \\ \sum_{A_i \cap B_j = C} \frac{m_1(A_i) \cdot m_2(B_j)}{1 - k}, & \end{cases}$$

where

$$k = \sum_{A_k \cap B_l = \phi} m_1(A_k) \cdot m_2(B_l).$$

The resultant  $m(C)$  is a new body of evidence representing the combination of two original bodies of evidence. The new evidence may in turn be combined with evidence from other sources. This is the process of belief of propagation in D-S theory (Garvey *et al* 1981).

### 3.4 Suitability to remote sensing

The suitability of the Dempster-Shafer theory for remote sensing is justified because of the following reasons.

- (1) The combination rule tries to discard conflicts by way of normalization and brings out consensus.
- (2) Order of combination is immaterial because of commutativity and associativity of multiplication which is the primitive operation in belief combination and propagation.

- (3) Dempster's combination rule acts over the entire subset space. Because of this, computations grow exponentially over the set of identification names i.e.,  $F$  (frame of discernment). But in remote sensing image interpretation, some subsets are not required to be taken into consideration, as for such subsets there will be no evidence since spectral signature alone is sufficient to label some objects.
- (4) Ignorance of the user in apportioning his belief can be carried out till the end of processing in a structured way. With this facility, the user is not forced to label his belief to any one or a combination of identification names.

### 3.5 Example of application of Dempster's combination rule

Suppose we have some possible subsets of identification names that are contributing evidence as indicated in the figure 4.

Evidence 1 from feature, colour:

- Belief in vegetation = 0.3,  
belief in soil = 0.5,  
not known (undistributed) = 0.2.

Evidence 2 from feature, texture:

- Belief in soil or water = 0.7,  
not known (undistributed) = 0.3.
- Summed up value for vegetation = 0.09,  
summed up value for soil or water = 0.14,  
summed up value for soil =  $0.35 + 0.15 = 0.5$ ,  
summed up value for undistributed = 0.06,  
conflict (null set) = 0.21,  
pooled belief for vegetation =  $0.09 / (1 - 0.21) = 0.11$ ,  
pooled belief for soil or water =  $0.14 / 0.79 = 0.18$ ,  
pooled belief for soil =  $0.59 / 0.79 = 0.63$ ,  
uncertainty =  $0.06 / 0.79 = 0.08$ ,  
plausibility(soil) =  $1 - \text{Bel}(\neg \text{soil}) = 1 - \text{Bel}(\text{vegetation}) = 1 - 0.11 = 0.89$ ,  
evidential interval for soil is  $[0.63, 0.89]$ ,  
ignorance = 0.26.

Interpretation of results:

- With the available evidence 'soil' is the identification name, considering maximum value of belief.

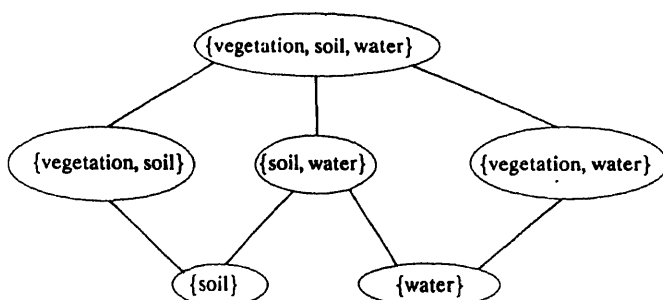


Figure 4. Sample subsets of a set of identification names contributing evidence.

#### 4. Methodology

We have bifurcated a given image into logical objects manually into the appropriate types such as point, line or region. Let  $s[i]$  be the set of possible identification names representative of  $i$ th feature value given by a user with his confidence value  $c[i]$ . Thus we have  $N$  sets of  $s[i]$  and  $c[i]$  for  $i = 1, 2, \dots, N$ . Let  $m(s[i])$  denote the weightage (or a portion of belief) indicating that the identification name is in the subset  $s[i]$  of  $F$ , frame of discernment. Assignment of this weightage is the crux of the problem and is the basis for getting successful results. Two methods are adopted to decide  $m(s[i])$ .

##### 4.1 Basic probability assignment

*Method 1:* User's confidence in the description of interpretation key value is taken directly as  $m(s[i])$ . This is analogous to the way an expert does, that is, totally depending on his confidence. With  $m(s[i])$ , weightage for a particular identification name is calculated by summing up the confidence values of sets in which the identification name occurs. Thus the name with maximum weightage is considered as the identification name. This method is termed "confidence".

*Method 2:* The idea of taking  $c[i]$  as  $m(s[i])$  directly may lead to a brittle decision because of the following. While inputting the confidence values, humans may not be consistent always. It is difficult to apportion belief in the same proportions always. In such situation, we feel that the cardinality of the subset  $s[i]$  is to be taken into consideration and we have done so in calculating  $m(s[i])$ . Thus  $m(s[i])$  is the product of  $c[i]$  and  $1/|s[i]|$ ; here equal chance of occurrence of any name in the set  $s[i]$  is the heuristic. This method is named "system". Dempster's combination rule is applied to these two methods.

##### 4.2 Interpretation rules

The Dempster-Shafer approach provides specific numerical values of belief and plausibility allowing the residual uncertainty to exist. Interpretation of these values as qualitative results is to be done by the system designer. We have interpreted the results depending on following interpretation rules.

- (1) Label the identification name having maximum plausibility and belief value compared to all others.
- (2) If two labels have the same belief, then the one with the higher plausibility is considered. This is because the same belief does not mean the same plausibility.
- (3) If two or more labels have the same belief and plausibility then suspend judgement and guide the user to go for collateral data.
- (4) Set threshold values for belief, plausibility and evidential interval and judge the label name.

##### 4.3 Knowledge organization and control sequence

Normally ordering of pattern features has a direct effect on the efficiency of recognition (Makato 1984). But in the D-S method it is immaterial because of associativity and commutativity of multiplication operation which is the key for combination and

propagation of evidence. Our method is the bottom-up procedure in which we constructed the required evidence from the feature values along with user's confidence. A sequential method of control is used for the identification of category name of a given object on FCC. The sequence used is pattern feature colour, pattern, texture, shape, size, association, relief and season. Each step is a partial decision making step, precipitating the available evidence to formulate subcategories. Steps go on until a subcategory contains only one identification name or no more evidence accumulation is possible.

## 5. Description of the system

The functional flow diagram of the software system is shown in figure 5. The software package is menu-driven having facilities to store facts and rules, to store image objects to be identified, to modify facts and rules and the inferencing mechanism to identify a target object. Explanation is provided at user's option and 'learning' is incorporated which uses its experience acquired based on the systems previous usage. Appropriate warnings and explanatory messages are given at the required places for an easy operation of the software. Summary of programs developed for construction of the knowledge bases and identification is given below.

### 5.1 Construction of knowledge bases

The knowledge base of the system consists of facts-base REGFAC.LSP, rule-base OBJRUL.LSP and learning sets LEARN.LSP occupying a storage space of 51 k bytes. The knowledge base can be updated and modified as and when the new facts are collected. From the interactive session with user, the system itself chooses and forms

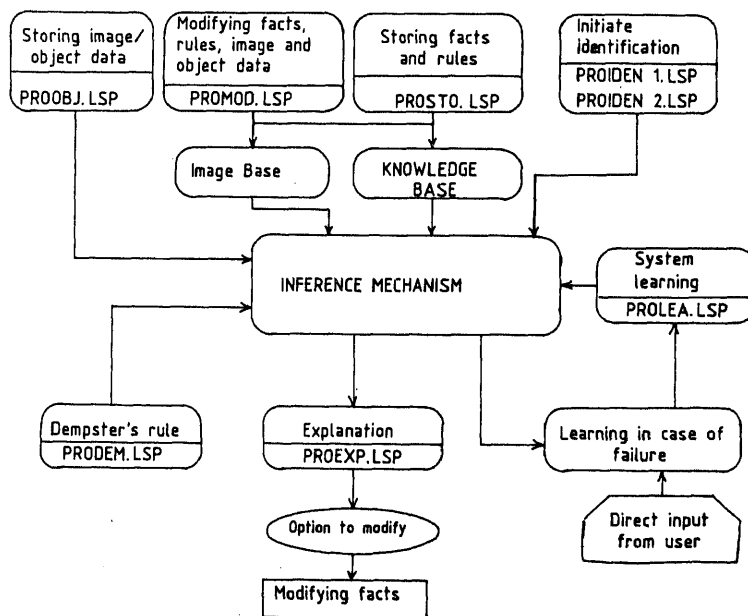


Figure 5. Functional flow diagram of the software.



appropriate knowledge structure and stores in the knowledge base. Some of the lists maintained by the system's knowledge base are given below.

(i) REGION/LINE/POINT

This consists of a set of possible identification names on which facts/rules exist in knowledge base. The universe of discourse or frame of discernment is formed here.

(ii) RULE

This consists of the set of all rule name,  $RULE_i$ , in the system. Also the facts in REGFAC.LSP are classified as class and subclass.

The data of logical image object to be identified on a given FCC is stored in the form of list in IMAGE.LSP. This maintains two lists namely, IMAGE containing FCC imagery  $I_i$  and OBJECT containing the set of all objects,  $O_i$ , pertaining to  $I_j$ .

## 5.2 Identification of image objects

The identification mechanism is initiated by the files PROIDEN1.LSP and PROIDEN2.LSP. These files have 26 functions constituting the inference mechanism for identifying a target object on FCC whose data is stored in IMAGE.LSP. Also they call functions in PRODEM.LSP, PROLEA.LSP and PROEXP.LSP according to the options exercised by user.

Once the object  $O_i$  on an FCC is chosen, based on the type of object, namely region or line or point, the respective expert, namely, REGEXP/LINEXP/POIEXP is triggered. These experts make use of facts and rules in knowledge-base (REGFAC.LSP and OBJRUL.LSP) and lead to formation of ten sets  $S_1$  to  $S_{10}$ , one for each feature, which contain the possible identification names based on the match between the corresponding feature value of the object to be identified and that of identification name. List of confidence values entered by the user for each feature of the object to be identified is formed in CONF. List of pairs (sub-expert, rules fired) is stored in RESUL1. Having formed the above mentioned lists and sets, the function INFER of PROIDEN2.LSP is executed, which gives the user the options of methods of identification of the object chosen as shown in figure 6. For DEMPSTER method, functions in file PRODEM.LSP are made use of. If the user wishes to use the learning done by the system previously, learning sets in LEARN.LSP are made use of.

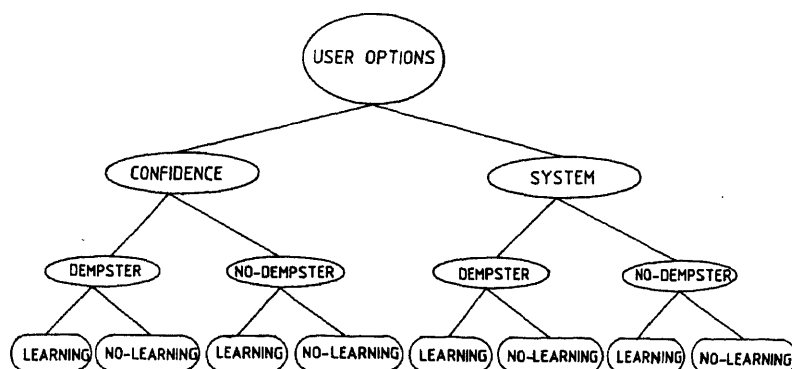


Figure 6. Options in the method of image-object identification.

## *Explanation*

These intermediate results and procedural explanation for each method of identification selected, can be seen during explanation session, which makes use of file XP.LSP. Explanation facility is provided by tracing the rules fired in the sequence giving the plausibility and belief values for the set in which the identified object member.

## *Learning*

When the result arrived is an incorrect one, the user can give the correct answer so the system can reallocate the weightages to each of the sets  $S_i$  during identification given at the correct solution. Sometimes, learning may fail if all the sets  $S_i$  containing correct answer, also contain the incorrect answer in which case the user is prompted an appropriate message.

## *Computation complexity*

Dempster's combination rule acts over the entire subset space of frame of discernment, the set of identification names. Hence identification by the Dempster-Shafer approach has computation complexity of the order  $O(k.n.2^n + c)$ , where  $n$  is the cardinality of frame of discernment and  $k$  and  $c$  are constants.

## **Results of identification**

Knowledge base of the system has been developed and tested on the basis of FCC of IRS-1A in addition to a hypothetical test image. In this paper we present details of one FCC (shown in figure 7) and the hypothetical test image. Some of detectable image-objects are indicated on the images by decimal numbers. The names of these objects are extracted upon consultation with a skilled interpreter.

The results of identification of the image-objects, namely O19 to O23, are shown in figure 8. Objects O1 to O4 belong to the hypothetical image I1 and are considered for purpose of testing. Objects O19 to O23 belong to the image shown in figure 7 and are tagged to a symbolic image name I4 in the knowledge-based system. Objects

O14 belong to a symbolic image I2 and objects O15 to O18 belong to I3. For session on the objects O5 to O18 the reader is referred to Sarma (1991). Results are obtained by exercising all the 8 options as shown in figure 6. Each user option is a path from the root to a leaf node. The results are compared with an expert's opinion as shown in the last column of figure 8. Results are correct to the extent of 95% in land use and cover domain (coastal belts) for which we have developed the knowledge base.

## *Critical evaluation*

From the results obtained it is observed that SYSTEM measure of obtaining  $m(S[i])$  is more accurate than CONFIDENCE measure. The D-S approach of identification with SYSTEM measure is more accurate, and it also helps in analysing the results with respect to plausibility, belief and evidential interval. It is appropriate to highlight a case of object O2, with options CONFIDENCE and NODEMPSTER, learning may

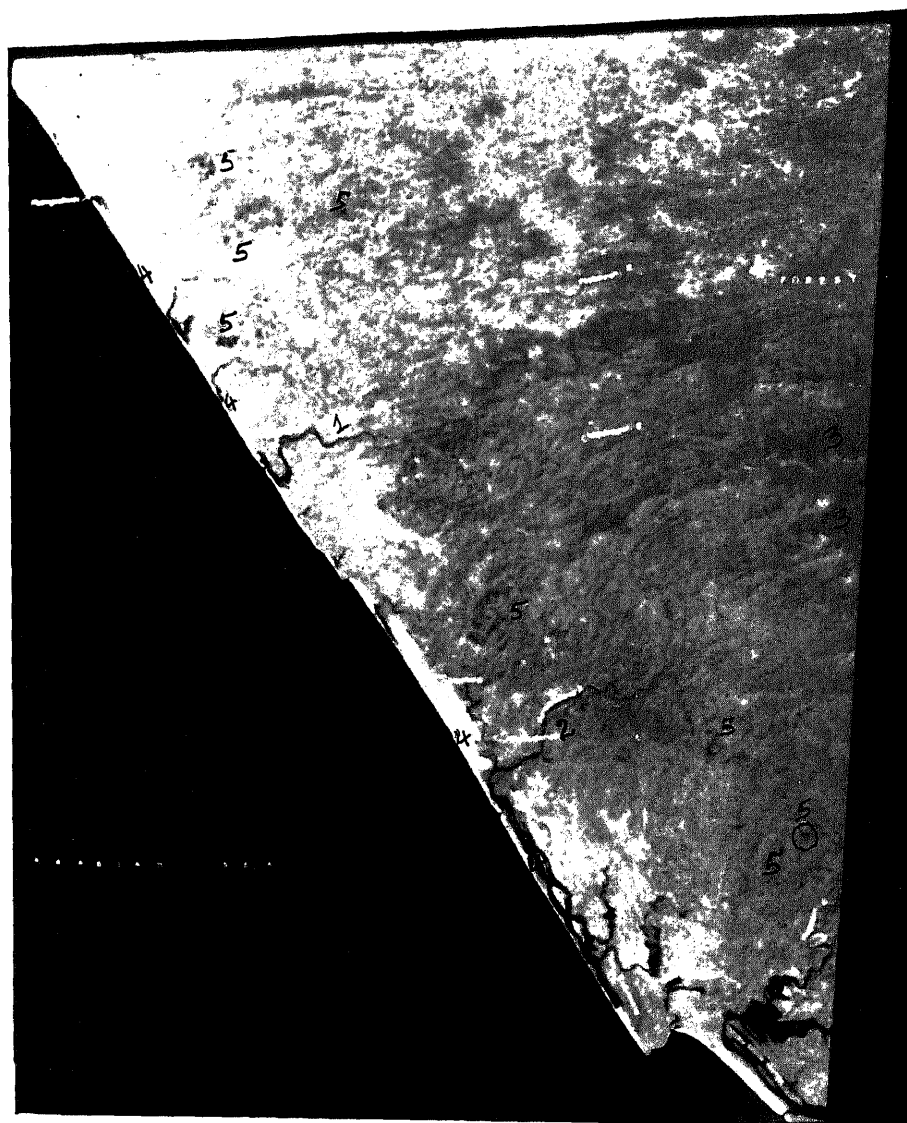


Figure 7. FCC under interpretation.

fail, assuming the correct identification is  $R3$  (fallow land). It is due to the fact that  $R3$  occurs in the set ( $R3, R14, R28$ ), and is an incorrect answer. So giving higher weight to sets  $S_i$  containing  $R3$  would ultimately result in increasing the weightage for the incorrect result of identification also.

It is evident that the result of identification of object  $O21$  with options CONFIDENCE and DEMPSTER with No-learning is wrong, that is,  $O21$  has been identified as crop land ( $R2$ ) instead of evergreen forest ( $R6$ ). Further the system has been provided with the correct answer  $R6$  during the learning session. The rock exposures (laterite cappings) on the image can be seen as point objects spread almost throughout the coastal side of the scene. These are identifiable because of the significant feature, the



5	012	LINE	River	River	River	River	River	River	River	River	River	River
9	013	LINE	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream
8	014	LINE	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream
Image 1 I3	015	REGION	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11/Man-grove	R11
4	016	REGION	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14/Water-logged-L	R14
2	017	LINE	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream	Stream
3	018	LINE	River	River	River	River	River	River	River	River	River	River
Image 1 I4	019	LINE	River	River	River	River	River	River	River	River	River	River
fig.7 2	020	LINE	River	River	River	River	River	River	River	River	River	River
3	021	REGION	R6/Forest evergreen	R2/cropland	R6/Forest evergreen	R6/Forest evergreen	R6/Forest evergreen	R6/Forest evergreen	R6/Forest evergreen	R6/Forest evergreen	R6/Forest evergreen	R6
4	022	REGION	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30/Coastal sands	R30
5	023	POINT	Rock exposures	Rock exposures	Rock exposures	Rock exposures	Rock exposures	Rock exposures	Rock exposures	Rock exposures	Rock exposures	Rock exposures

Figure 8. Results of identification.

relief of undulating low hills devoid of vegetation. From the photograph in figure 7, the relief may not be striking. However, in the original image it is apparent.

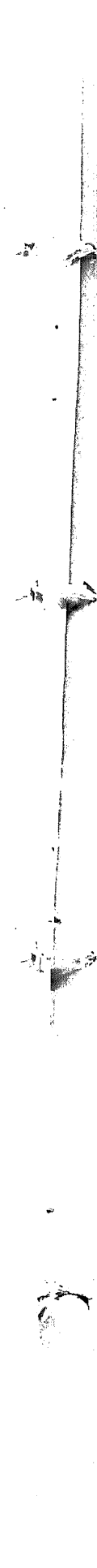
## 7. Conclusions

In this paper, we have demonstrated the application of knowledge-based methods for remote sensing satellite image interpretation. Our motivating assumption, that image interpretation is a form of intelligence-computation involving qualitative reasoning, is realized in the process of development of the prototype expert system. We have considered two basic probability assignment methods, namely 'confidence' and 'system' and combined each one with (or without) Dempster's combination rule (with or without learning). Thus 8 options have arisen for identification. We have carried out the identification process exercising all these options in a bid to analyse the consistency and correctness of the methods and found that the 'system' method is more accurate than the 'confidence' method. For image objects which have similar features, threshold values for plausibility, belief and evidential intervals are critical for correct identification. It is observed that the time taken to reason with regard to objects represented in property lists is significantly large. So we have decided to maintain the knowledge of line and point objects in the form of rules, leaving region objects' data in property lists. This has improved the speed of execution though the overall computation complexity is same.

Although we have taken the standard FCC of IRS-1A for identification and analysis of results, use of the system is not restricted to FCC only. It can also be used to interpret black-and-white images or any other photographic data products for which experts can design an interpretation key. Since the system has taken the shape of an expert system shell, removing the existing knowledge base and providing a new knowledge base would enable it to be used for the interpretation tasks described above. At present, our system may be used as an aid to an expert interpreter.

## References

- Argialas P D 1990 Computational image interpretation models: An overview and perspective. *Photogrammetric Eng. Remote Sensing* 56(6): 871-876
- Garvey T D, Lowrance J D, Fischler M A 1981 An inference technique for integrating knowledge from disparate sources. *International Joint Conference on Artificial Intelligence 1981 Proceeding*: 319-325
- Hayes-Roth F 1989 Towards benchmarks for knowledge systems and their implications for data engineering. *IEEE Trans. Knowledge Data Eng.* 1(1): 101-110
- Makato N 1984 Control strategies in pattern analysis. *Pattern Recogn.* 17(1): 45-56
- Ng K C, Abramson B 1990 Uncertainty management in expert systems. *IEEE Expert*: 29-47, April 1990
- Sarma L C 1991 *Knowledge based interpretation of remote sensing data*. M Sc (Eng.) Thesis, Department of Computer Science and Automation, Indian Institute of Science, Bangalore
- Sarma L C, Sarma V V S 1990 A knowledge based system for land use/land cover categorization. *Proceedings of Artificial Intelligence and Expert System Technologies in the Indian Context*, Tata McGraw-Hill Publishing Company Ltd., Volume 2: 108-114
- Schowengerdt R, Wang H 1989 A general purpose expert system for image processing. *Photogrammetric Eng. Remote Sensing* 55: 1277-1284
- Wang F, Newkirk R 1988 A knowledge based system for highway network extraction. *IEEE Trans. Geosci. Earth Sci.* 26(5): 525-531
- Wharton W S 1987 A spectral knowledge-based approach for urban land cover discrimination. *IEEE Trans. Geo-Sci. Remote Sensing* 25: 272-282



## ***Paninian framework and its application to Anusaraka***

AKSHAR BHARATI, VINEET CHAITANYA and RAJEEV SANGAL<sup>1</sup>

Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur 208 016, India

<sup>1</sup>E-mail: sangal@iitk.ernet.in

**Abstract.** The *Paninian*\* framework proposes *karakas* as semantico-syntactic relations that play a crucial role in mediating between surface form and meaning. The framework accounts for theta-role assignment, active passive, and control in a uniform manner. It has been successfully used in building an extremely fast prototype machine-translation system between two Indian languages. The constraint parser and the generator are designed with information theoretic considerations. Paninian framework is particularly suited to free word order languages. As most human languages are relatively word-order free, the Paninian framework should be explored as a serious contender for such languages. Based on the Paninian theory, the concept of language accessor or *anusaraka* has emerged, which has the potential to overcome the language barrier in India.

**Keywords.** *Paninian* grammar; *karakas*; *anusaraka*; machine translation; language accessor.

### **1. Introduction**

How is it that natural language is used by speakers to convey information to the hearers? How is it that on hearing an utterance, the hearers are able to get intended information? These are the questions that have intrigued *Paninians*. The goal of the Paninian approach is to construct a theory of human natural language communication that answers these questions. Grammar, a part of such a theory of communication, is a system of rules that establishes a relation between what the speaker decides to say and his utterance, and similarly, what the hearer hears and the meaning he extracts<sup>1</sup>.

The main problem that the Paninian approach addresses is how to extract *karakas* relations (which are syntactico-semantic relations) from a sentence. As it is inspired

---

\*In this paper non-English words occur many times and hence are italicized only at first mention.

<sup>1</sup>There is a difference between the goals of the Paninian approach on one hand and generative enterprise on the other. The latter is interested in identifying the innate universal grammar in the mind of every human child by which it so effortlessly and without explicit tutoring acquires language that it is exposed to.



by an inflectionally rich language, it emphasizes the roles of case endings and markers such as post-positions (or prepositions). Position or word order is brought into consideration only when necessary.

In the next section, we pose two problems in Hindi, a 'free' word order language: (1) how to identify semantic relations (or *karaka* role) from *vibhakti* markers, and (2) how to identify word senses. In § 3, we show how the Paninian theory accounts for the problems. Section 4 discusses computational aspects.

A majority of human languages including Indian and other languages have relatively free word order. In free word order languages, the order of words contains only secondary information such as emphasis etc. Primary information relating to 'gross' meaning (e.g., one that includes semantic relationships) is contained elsewhere. In contrast to these languages, most existing natural language processing systems are based on context free grammars which are basically positional grammars. These are designed for languages in which the position of a constituent in a sentence contains key information. It is important to develop a suitable computational grammar formalism for free word order languages for two reasons:

- (1) a suitably designed formalism will be more efficient because it will be able to make use of primary sources of information directly;
- (2) such a formalism is also likely to be linguistically more elegant and satisfying. Since it will be able to relate to primary sources of information, the grammar is likely to be more economical and easier to write.

In this paper, we describe such a formalism, called the Paninian framework, that has been successfully applied to Indian languages. The parser and the generator have been designed for the framework based on information theoretic considerations. These are part of a small prototype machine-translation system from Hindi to Telugu. The resulting system is extremely fast, because the grammar makes direct use of *vibhakti*, the source of primary information in Indian languages, and the parser makes efficient use of grammar while parsing.

## 2. Indian languages: Some salient features

### 2.1 Free word order and *vibhakti*

Indian languages have relatively free word order. Many of the constituents of a simple sentence can occur in any order without affecting the gross meaning of the sentence; what is affected is perhaps the emphasis etc. For instance, noun groups in a sentence can come in any order without affecting the theta relationships (or semantic relationship between the verb and the noun groups). Since position or order of occurrence of a noun group does not contain information about the theta roles in a simple sentence, a question can be asked regarding what carries this information. The answer seems to be that post-position markers after nouns (in North Indian languages) or surface case endings of nouns (in South Indian languages) or a mixture of the two at times, play a key role in specifying semantic relationships. We will collectively refer to the post position markers and surface case endings of nouns as *vibhakti*. Consider the following sentences in Hindi (Bharati et al 1990a, 1991a).

- A.1 *ra:m mo:han ko: piṭṭa: hai*  
 Ram Mohan -ko: beat is.  
 'Ram beats Mohan'.

A.2 *mo:han ko: ra:m piṭṭa: hai*

Mohan-ko: Ram beat is.

'Ram beats Mohan'.

A.3 *mo:han ra:m ko: piṭṭa: hai*

Mohan Ram -ko: beat is.

'Mohan beats Ram'.

A.4 *ra:m ko: mo:han piṭṭa: hai*

Ram -ko: Mohan beat is.

'Mohan beats Ram'.

In A.1 and A.2, Mohan has the same vibhakti (i.e., *parsarg* or post position 'ko:') and semantic relation with beating. Even though the position of 'ra:m' and 'mo:han ko:' are interchanged in A.2, it does not alter the respective semantic relations of Ram and Mohan with the verb. A.3 and A.4 show that semantic relation of Ram is interchanged with that of Mohan by interchanging their vibhakti. So the vibhaktis are crucial in determining the semantic roles. The relative position of the nominal does not seem to be very important for determining semantic relations.

However things are not always straightforward and the following need to be accounted for: A different vibhakti can be used for the same semantic relation with a given verb in a different sentence. For example, in the Hindi sentence B.1 to B.4, although Ram has the same semantic relation with eat (namely, the agent of eat), a different vibhakti is used each time (nil, *ne:*, *ko:*, *se:*, respectively).

B.1 *ra:m p<sup>h</sup>al ko: k<sup>h</sup>a:ta: hai:*

Ram fruit -ko: eats is.

'Ram eats the fruit'.

B.2 *ra:m ne: p<sup>h</sup>al k<sup>h</sup>a:ya:*

Ram -ne: fruit ate.

'Ram ate the fruit'.

B.3 *ra:m ko: p<sup>h</sup>al k<sup>h</sup>a:na: pada:*

Ram -ko: fruit eat had to.

'Ram had to eat the fruit'.

B.4 *ra:m se: p<sup>h</sup>al nahi: k<sup>h</sup>a:ya: gaya:*

Ram -se: fruit not eat could.

'Ram could not eat the fruit'.

## 2.2 Head and vibhakti

Vibhakti for nouns has already been defined earlier. A noun group is a unit containing a noun (or a pronoun, proper name etc.), its vibhakti and possibly the same objectives. The noun is the head and it occurs close to the lexical items that express its vibhakti. In case, the vibhakti is expressed by means of surface case ending, it is incorporated in the head word by morphological process; and in case, it is expressed by means of a post-position marker or *parsarg* and normally occurs immediately after the head. Sometimes a particle can intervene between the head word and the *parsarg*, but the *parsarg* is within a bounded distance (usually just one or two words away) from the head. This property has an important implication for parsing strategy: the vibhakti of a nominal can be identified first even before identifying the noun group (for example, before identifying adjectives or other modifiers).

Vibhakti for verbs can be defined similar to that for the nouns. A head verb may be followed by auxiliary verbs (which may remain as separate words or may combine with the head verb). Such information is collectively called vibhakti for the verb. The vibhakti for verb gives information about tense, aspect and modality (TAM), and is, therefore, also called the TAM label. TAM labels are purely syntactic determined from the verb form and the auxiliary verbs<sup>2</sup>.

A verb group consists of the head verb, its vibhakti, and possibly some particles for emphasis, negation etc. Again, the head verb occurs close to the lexical items that express its vibhakti.

### 3. Paninian theory

Paninian grammar is particularly suited to free word order languages. It makes use of vibhakti information for mapping to semantic relations, and uses position information only secondarily. As the Indian languages have (relatively) free word order and vibhakti, they are eminently suited for description by the Paninian grammar. The Paninian framework was originally designed more than two millenia ago for writing a grammar for Sanskrit; it has been adapted by us to deal with modern Indian languages.

#### 3.1 Karaka relations

Example sentences in Hindi from the previous section (in A and B) indicate that there is no straightforward mapping from vibhakti to semantic relations between noun groups and verbs. The key to arriving at an answer is to identify intermediate relations.

The notion of *karaka* (pronounced 'ka:ra:k') relation is central to the model. These are semantico-syntactic relations between the verb(s) and the nominals in a sentence. The computational grammar specifies a mapping from the vibhaktis of nominals and the verb(s) in a sentence to karaka relations between them. Similarly, other rules of grammar provide a mapping from karaka relations to semantic relations between the verb(s) and the nominals. Thus, the karaka relations by themselves do not give the semantics. They specify relations which mediate between vibhakti of nominals and verb forms on the one hand and semantic relations on the other (Kiparsky 1982). Figure 1 shows the relationship pictorially.

The *karta* karaka holds between that nominal and a verb in a sentence, whose referent is 'swatantra' or the most independent or autonomous out of all the karaka

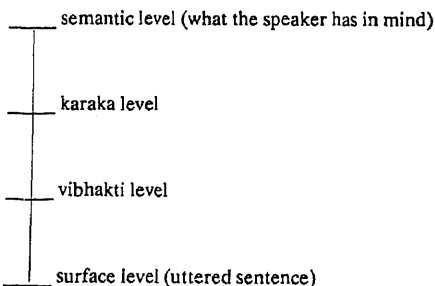


Figure 1. Levels in the Paninian model.

<sup>2</sup> A suitable mapping, though complex, can take us from TAM labels to TAMs.

Karaka	Vibhakti	Presence
Karta	Ø	mandatory
Karma	ko: or Ø	mandatory

Figure 2. Default mapping for some karakas.

nominals that are expressed by the speaker. However, it is with respect to the activity implied by the verb.

### 3.2 Karaka to vibhakti mapping

We are now ready to discuss the mapping from karaka level to vibhakti level. This will allow us to map a representation of a sentence at the karaka level to a representation at the vibhakti level (or the inverse). We will not have anything further to say, in this paper, on the mapping from semantic level to karaka level.

The most important insight regarding the karaka-vibhakti mapping is that it depends on the TAM label of the verb. There is a default mapping from karakas to vibhaktis of nominals in a sentence. The default mapping holds if the verb in question (whose karaka relations are given) has a particular TAM label called basic. If the verb has some other TAM label, the default is transformed depending on the TAM label<sup>3</sup>.

In Hindi for instance, the basic TAM label is *ta: hai* (which roughly stands for the present indefinite). The default mapping for two of the karakas is given in figure 2. This explains the vibhaktis in sentences A.1 to A.4. As Ram is the agent in A.1 and A.2, and Mohan in A.3 and A.4 and the agent is the most independent for the action 'beat', it is expressed by means of the karta karaka; and the remaining nominal by *karma* karaka. For the karta karaka, 0 vibhakti is used with *ta: hai* TAM label in A.1 to A.4 as explained in figure 2.

Figure 3 gives some transformation rules for the default mapping for Hindi. It explains the vibhakti in sentences B.1 to B.4 (assuming that Ram is the karta and *p<sup>h</sup>al* 'fruit' is the karma). As explained by figure 3, karta takes 'ne:' vibhakti in B.2 because of TAM label 'ya:' (in the main verb group *k<sup>h</sup>a:ya:*), 'ko:' in B.3 because of TAM label *na: pada:* (in *k<sup>h</sup>a:na: pada:*), and 'se:' in B.4 because of TAM label *ya: gaya:* (in *k<sup>h</sup>a:ya gaya:*).

### 3.3 Control

A major support for the theory comes from complex sentences, that is sentences containing more than one verb group. We first introduce the problem and then describe how the theory provides an answer. Consider the following sentences in Hindi.

- G.1 *ra:m p<sup>h</sup>al k<sup>h</sup>a:kar mo:han ko: bula:ta: hai:*  
 Ram fruit having-eaten Mohan -ko: calls is  
 'Having eaten fruit, Ram calls Mohan'

TAM label	transformed vibhakti for karta
<i>ya:</i>	<i>ne:</i>
<i>na: pada:</i>	<i>ko:</i>
<i>ya: gaya:</i>	<i>se:</i> or <i>dwa:ra:</i> (and karta is optional)

Figure 3. Transformation rules.

<sup>3</sup>What karaka relations are permissible for a verb, obviously, depend on the particular verb. Not all verbs will take all possible karaka relations.

TAM label	transformation
<i>kar</i>	Karta must not be present Karma is optional
<i>na:</i>	Karta and karma are optional
<i>ta:_hua:</i>	Karta and karma are optional

Figure 4. More transformation rules.

G.2 *ra:m ne: p<sup>h</sup>al ka:ṭkar k<sup>h</sup>a:ya:*

Ram *ne:* fruit having-cut ate

'Ram ate having cut the fruit'

G.3 *p<sup>h</sup>al ka:tne: ke: liye: usne: ca:ku: liya:*

fruit to-cut -*ke:-liye:* he:-*ne:* knife took

'To cut fruit, he took a knife'

In G.1, Ram is the karta of both the verbs: *k<sup>h</sup>a:* 'eat' and *bula:* 'call'. However, it occurs only once. The problem is to identify which verb will control its vibhakti. In G.2, karta Ram and the karma *p<sup>h</sup>al* 'fruit' both are shared by the two verbs *ka:t* 'cut' and *k<sup>h</sup>a:* 'eat'. In G.3, the karta *usne:* 'he' is shared between the two verbs, and '*ca:ku:*' 'knife' the karma karaka of *le:* 'take' is the karana (instrumental) karaka of '*ka:t* 'cut'.

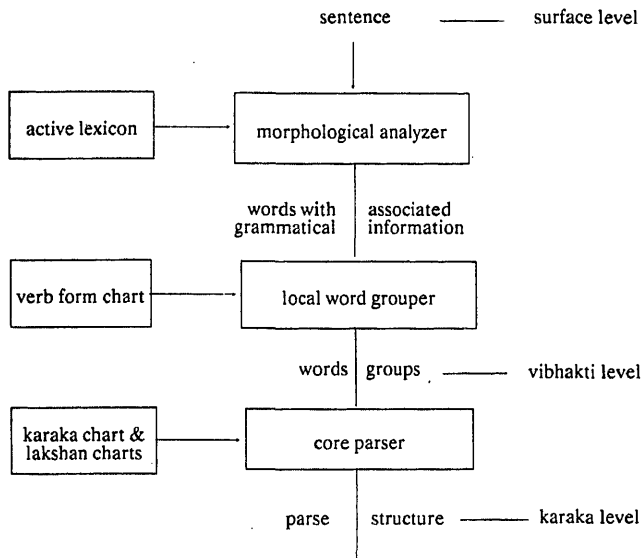
The observation that the matrix verb rather than the intermediate verb controls the vibhakti of the shared nominal is true in the above sentences. The theory we will outline to elaborate on this theme will have two parts. The first part gives the karaka to vibhakti mapping as usual, the second part of the theory (not described in this paper) identifies shared karakas. See Bharati *et al* (1994) for details.

The intermediate verbs have their TAM labels just like other verbs. For example, *kara* is the TAM label of *k<sup>h</sup>a:* 'eat' in G.1, and the *na:* is the TAM label of *ka:t* 'cut' in G.3. As usual, these TAM labels have transformation rules that operate and modify the default karaka to vibhakti mapping. In particular, the suggested transformation rules for the two labels are given in figure 4. The transformation rule with *kara* in figure 4 says that karta of the verb with TAM label *kara* must not be present in the sentence and the karma is optionally present.

By these rules, the intermediate verb *k<sup>h</sup>a:* 'eat' in G.1 does not have (independent) karta karaka present in the sentence. Ram is the karta of the matrix or the main verb *bula:* 'call'. *p<sup>h</sup>al* 'fruit' is the karma of *k<sup>h</sup>a:*. All these are accommodated by the above transformation rule for '*kara*'. The karta of *k<sup>h</sup>a:* in G.1 and similarly in other cases, can be obtained by sharing rules, which are not described here.

#### 4. Information-theoretic parser

The Paninian theory outlined above can be used for building a parser. (See Bharati *et al* (1990b) for a discussion on other approaches to parsing.) Parsing is the reverse of generation, where given a sentence a suitable semantic structure is to be assigned to it. If we build a parser based on the Paninian theory, we have to obtain a representation of a given sentence at the vibhakti level, using which we must obtain a representation at the karaka level, and finally, a representation at the semantic (or mental) level. (Again, in this paper we will not talk about the semantic level, and will focus on mapping from the vibhakti level to the karaka level.)



**Figure 5.** Structure of the parser.

It turns out that the Paninian theory is extremely suitable from the computational viewpoint. It can be used in a natural manner for structuring a parser which is extremely efficient.

It is fairly obvious that one part of the parser must take care of morphology. For each word in the input sentence, a dictionary or a lexicon needs to be looked up, and associated grammatical information needs to be retrieved. The words have to be grouped together yielding nominals, verbals etc. Finally, the karaka relations among the elements have to be identified. This is shown in figure 5.

#### 4.1 Information-theoretic approach

The parser is based on information-theoretic considerations where at each stage of processing, just the right amount of information is extracted. At the morphological analysis stage, information available from word forms is obtained. For example, information about gender, number, person is obtained (wherever possible) from the nouns. On the other hand, for verbs in Hindi, gender, number, person (*gnp*) and part of TAM is obtained from the words. In case of Telugu verbs the complete TAM label, besides *gnp* is obtained as well.

In the local word grouping stage, words combine into groups (noun groups and verb groups) based on local information. The groups are formed with minimal computational effort (finite-state machine model) as only local information is used. On the other hand, local word grouping brings together just the right information (vibhakti for nouns, TAM labels for verbs) that is needed for the next stage of processing (that is, for karaka assignment). It does not attempt to distinguish all the fine shades of meaning, for example, temporal structure or modality. For one thing, such information cannot easily be determined at this stage of processing. Secondly, it is not needed for the next stage.

As mentioned earlier, the output of the local word grouper roughly corresponds to the vibhakti level. However, in those few cases where there is ambiguity in identifying the local word groups, which cannot be resolved at that level, the decision

is postponed. For example, in Hindi, a word that can be both a noun and an adjective, causes ambiguity in forming a local word group with its succeeding noun. The choice can only be made later during karaka assignment using karaka charts. As a result, in our information-theoretic parser, such a choice is delayed to the point when it can be made. Similar is the case with a noun that is followed by a marked *ka:*, *ke:* or *ki:* and succeeded by a noun.

After the local word grouping stage, there is karaka assignment and lexical disambiguation stage. This is done at this stage because the necessary information (vibhakti) for doing the above is available. Other phenomena such as quantifiers and anaphora are not handled because information for resolving them is not available.

This approach is consistent with the Indian grammatical analysis where meaning is extracted in several layers with increasing precision.

## 4.2 Core parser

Morphological analyser and local word grouper shown in figure 5, have been described elsewhere (Bhanumathi 1989; Bharati et al 1991b). Here we discuss the core parser.

Given the local word groups in a sentence, the task of the core parser is two fold:

- (1) to identify karaka relations among word groups, and
- (2) to identify senses of words.

The first task requires knowledge of karaka-vibhakti mapping, optionality of karakas, and transformation rules. The second task requires *lakshan* charts for nouns and verbs, to be discussed later.

A structure called karaka chart stores information about karaka-vibhakti mapping and optionality of karakas. Initially, the default mapping is loaded into it for a given verb group in the sentence. Transformations are performed using the TAM label. There is a separate karaka chart for each verb group in the sentence being processed. Information about semantic types of fillers of karaka roles is also available. But such information is limited to that necessary for removing ambiguity, if any, in karaka assignment. In other words, for a given verb, when karaka-vibhakti mapping is not sufficient for producing an unambiguous parse, semantic types are included. The semantic types so included have the sole-purpose of karaka diambiguation. This keeps the number of semantic types under control, and serves as a guiding philosophy for what semantic types to include. Figure 6 shows the starting semantic type hierarchy which is sufficient for a major part of language. An example karaka chart for *k<sup>h</sup>a:* 'eat' is given in figure 7. It shows for each of the karakas its necessity (mandatory, desirable, or optional), vibhakti, and semantic type. These are called karaka restrictions in a karaka chart.

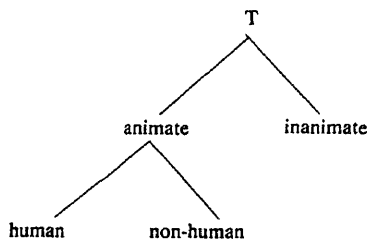


Figure 6. Semantic type hierarchy.

karaka	necessity	vibhakti	semantic type
karta	m (mandatory)	Ø	animate
karma	m	Ø - or - ko:	--

Figure 7. Karaka chart for  $k^ha$ .

For a given sentence after the word groups have been formed, karaka charts for the verb groups are created and each of the noun groups is tested against the karaka restrictions in each karaka chart (provided the noun group is to the left of the verb group whose karaka chart is being tested). When testing a noun group against a karaka restriction of a verb group, vibhakti information and semantic type are checked, and if found satisfactory, the noun group becomes a candidate for the karaka of the verb group. This can be shown in the form of a constraint graph. Nodes of the graph are the word groups and there is an arc from a verb group to a noun group labelled by a karaka, if the noun group satisfies the karaka restriction in the karaka chart of the verb group. (There is an arc from one verb group to another, if the karaka chart of the former has a karaka restriction with semantic type as action). The verb groups are called demand groups as they make demands about their karakas, and the noun groups are called source groups because they satisfy demands. (A verb group can be a source group as well when it satisfies the demand of another verb group. This however does not affect its status as a demand group as well.)

As an example, consider a sentence containing the verb  $k^ha$ : 'eat' with its word groups marked:

*bacca: ke:le: ko: k^ha:ta: hai:*  
 child banana -ko: eats  
 'The child eats the banana'.

Its constraint graph is shown in figure 8. It also happens to be the solution graph. Consider another sentence where the constraint graph is different from the solution graph.

*bacca: ke:la: k^ha:ta: hai:*  
 child banana eats  
 'The child eats a banana'.

Here, both the nouns qualify, to be karta and karma. In such a situation, the parser produces both parses; however, the first parse is one in which the animate entity, namely *bacca:* (child), is the karta.

4.2a *Constraints:* A parse is a sub-graph of the constraint graph containing all the nodes and satisfying the following conditions (Bharati & Sangal 1990):

- (1) for each of the mandatory karakas in karaka chart for each demand group, there should be exactly one outgoing edge from the demand group labelled by the karaka;
- (2) for each of the desirable or optional karakas in a karaka chart for each demand group, there should be at most one outgoing edge from the demand group labelled by the karaka;

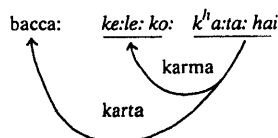


Figure 8. Constraint graph and solution graph.



- (3) there should be exactly one incoming arc into each source group.

If there are several sub-graphs for a constraint graph satisfying the above conditions, it means that there are multiple parses and the sentence is ambiguous. If no sub-graph satisfies the above constraints, the sentence does not have a parse.

**4.2b Constraint parser:** Currently, a parse is obtained from the constraint graph using integer programming. A constraint graph is converted into an integer programming problem by introducing a variable  $x_{ijk}$  for an arc from node  $i$  to  $j$  labelled by karaka  $k$  in the constraint graph such that for every arc there is a variable. The variables take their values as 0 or 1. A parse is an assignment of 1 to those variables whose corresponding arcs are in the parse sub-graph, and 0 to those that are not. Equality and inequality constraints in integer programming problem can be obtained from the conditions listed earlier, as follows, respectively:

- (1) for each demand group  $i$ , for each of its mandatory karakas  $k$ , the following equalities must hold:

$$\sum_j x_{ijk} = 1;$$

- (2) for each demand group  $i$ , for each of its optional or desirable karakas  $k$ , the following inequalities must hold:

$$\sum_j x_{ijk} < 1;$$

- (3) for each of the source groups  $j$ , the following equalities must hold:

$$\sum_{ik} x_{ijk} = 1.$$

The cost function to be minimized may be given as the sum of all the variables, in which case it does not show any preference for any of the parses.

It is possible, however, to put in preferences by suitably varying the cost function. If we use such a cost function, we will get a parse that has a minimum cost with respect to the cost function. The integer programming system can be so setup that we can ask for the next solution, in which case, we will get another parse with the same or higher cost. This can be repeated to obtain all the possible parses.

Some reasonable preferences which can be incorporated in the cost function are as follows: All else being equal,

- (1) karta has the following preferences in descending order: human, non-human, inanimate (animacy preference);
- (2) a source group is close to the demand group with which it has a relationship (closeness preference);
- (3) karta occurs before karma in a sentence (leftness preference)<sup>4</sup>.

<sup>4</sup>To the above list can be added a host of conditions dealing with anaphora, movement, garden-path sentences etc.

4.2c *Lakshan charts for sense disambiguation*: The second major task to be accomplished by the core parser is disambiguation of word senses. This requires the preparation of *lakshan* charts (or discrimination nets) for nouns and verbs.

A *lakshan* chart for a verb allows us to identify the sense of the verb in a sentence given its parse. *Lakshan* charts make use of the *karakas* of the verb in the sentence, for determining the verb sense. For example, in the sentence H.1 and H.2:

- H.1 *kisa:n k<sup>h</sup>e:t ko: jo:ṭta hai:*  
 farmer land -ko: ploughs  
 "The farmer ploughs the land"
- H.2 *kisa:n ga:di: ko: jo:ṭta: hai:*  
 farmer cart -ko: attaches  
 "The farmer attaches the cart"
- H.3 *kisa:n k<sup>h</sup>e:t ko: ka:ṭta: hai:*  
 farmer crops -ko: harvests  
 "The farmer harvests the crops"

The verb '*jo:ṭta: hai:*' is used in two different senses: plough or attach. *Lakshan* chart for '*jo:ṭ*' would allow us to select the appropriate sense of *jo:ṭ* by testing whether its *karma* is land or cart etc. Again, it is designed from an information theoretic point of view. The available information is used in an economical and efficient manner in deducing the right amount of new information.

A verb *lakshan* chart for a verb is prepared by linguists and language experts by looking up different senses of the verb with the help of conventional dictionaries. The chart builder must select features carefully that would allow verb sense to be obtained.

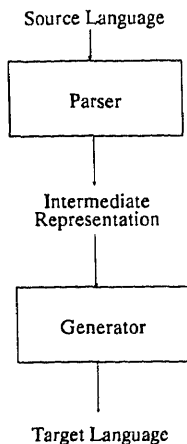
Noun *lakshan* charts help disambiguate senses of nouns in a sentence. They make use of the parse structure (i.e., *karka* relations) and the verb sense. For example, in sentences H.1 and H.3, the sense of *k<sup>h</sup>e:t* is land and crop, respectively. However, depending on the verb of which it is a *karma* the appropriate sense is selected.

Preparation of *lakshan* charts is a laborious process but is essential for fully-automatic high-quality machine translation. Linguistic theories are generally silent on the issue of word sense disambiguation.

## 5. An application: *Anusaraka* – a translation aid

The above theory can be used for building translation systems among languages. For translating from a source language to a target language, we need a parser for the source language and a generator for the target language (see figure 9). This is the interlingua approach. The choice of intermediate representation is important. For a machine translation (MT) system based on the Paninian theory, the intermediate representation is in terms of verb and noun concepts (i.e. word senses) and *karka* relations. For Indian languages, the quality of translation is fairly good using *karka* relations. Further analysis can improve the quality of translation.

Preparation of language data particularly *karka* charts and *lakshan* charts is a major task in building an MT system. It requires several man-years of effort per language. It turns out that one can talk about the concept of *anusaraka*.



**Figure 9.** A machine translation system.

### 5.1 What is language accessor

Anusaraka<sup>5</sup> or Language Accessor (LA) allows a reader who knows one language to access or follow text in another language. It produces output which is comprehensible to the reader, without worrying about it grammatically. LA is the result of a new way of looking at the problem: Instead of considering the problem as that of translation, it views the problem to be that of accessing information in another language with possibly some effort by the reader. Such a system for Indian languages, for example, does “simple” analysis (typically upto vibhakti level) of the source language sentences, and then generates sentences in the target language. The generated sentence need not be grammatical, however; it must be such that the reader can follow the meaning of the source sentence or text with little or some effort. Because of the relatively simple analysis done by the system, it is extremely robust towards improper input: ungrammatical sentences, sentence fragments, texts with ellipses, and other real life texts<sup>6</sup>. See Narayana (1994) for technical details.

### 5.2 Why LA

LA addresses several problems at once.

- (1) It makes a practical system available today without waiting for several years. Based on present dictionaries and other linguistic resources, it provides access to languages: a person can follow the meaning of a text in another language with some effort.
- (2) The work done in building LA will not go waste after high quality fully automatic machine translation (MT) systems are built. In fact, this work forms an essential

<sup>5</sup> *Anusarak* in Hindi literally means “follower”. It is different from “*anuvadak*” or translator. Anusarak allows a Hindi reader, for example, to read Hindi following Kannada or “Kannada *anusaar* Hindi”.

<sup>6</sup> It can be argued that there is no single “high quality” translation of a document. Quality or rather suitability of a translation depends on the purpose and the audience. For example, good translation of a document would be different for children and adults. Availability of Machine Translation or Language Accessor software and personal computers makes it possible to tailor the translator to the reader.

part of the future systems. Since the LA model is a part of the overall model for MT system, it will continue to develop and improve incrementally. It will not be a "dead-end" system.

- (3) It will function as an important tool for the linguist. The LA system will assist in the development of computational lexicons because it will throw up problem cases for comparative analysis.

Thus, LA delivers something practical today; the work to be done for building it, needs to be done in any case for MT; and its availability will help in the remaining work towards development of MT systems of the future.

LA brings about several changes in goals and design criteria when compared with conventional MT systems. Grammaticality reduces in importance, its place is taken by comprehensibility. It emphasizes robustness. (The system should never fail on any input!) It also assumes the availability of a personal computer to the person reading the output (unlike MT systems which assume that the translator uses the system to produce output for end users). It emphasizes the notion that multiple translations might be produced depending on the audience etc.

While LA and conventional MT systems might seem different, they are complementary. A practical system would also have an MT system besides the LA system in the background. Whenever possible MT would be tried, and in case of failure of the MT system to analyse the input etc., LA would be used to provide an answer. In this sense a practical LA system would keep improving as computational lexicons are prepared and MT systems become available.

### 5.3 Applications

LA can be seen as a core technology which can be put to varied uses. Like the internal combustion engine which can draw water, or move a car, or fly an aeroplane depending on how it is used, LA can also be used for different applications. It can assist a translator in the task of translation, provide direct access to books and printed matter to a reader in another language, provide access to documents in a multilingual setting within an organization such as the government, tourist assistance office etc.

LA can also be viewed as a telescope for the linguist. Just as the telescope allows the astronomer to see farther and more clearly, LA allows the linguist to observe phenomena in different languages at first-hand, rapidly and easily. It allows generation of larger masses of data and rapid testing of new theories. Sometimes users are able to get the flavour of the original sentence by looking at its "literal" translation. LA can provide such a flavour.

### 5.4 Some long term impact on education

Several long term impacts are obvious from the applications outlined above. The biggest impact would be the breaking of language barriers using LA. Here we discuss the possible impact on education and languages.

LA research is sharply focussed on those aspects of languages (in particular, of Indian languages) by which essential communication is possible. Thus, it tries to use those features of a set of languages which are readily transferable (with minimal analysis). It also identifies features which do not transfer well and either avoids using them or provides help when they occur. This research can lead to the identification

of a core which is common across Indian languages. This core can be taught in schools, which will allow people to pick any Indian language with ease. When LA systems proliferate, they will encourage learning of this core (as well as of the differences between particular languages that an individual user uses). It may turn out that with the widespread learning of the core, LA promotes development of languages closer to each other.

### 5.5 *Tasks in developing LA*

Development of LA requires joint work by linguists and language experts, and computer scientists. The former two will work on preparation of suitable electronic dictionaries for LA (based on existing dictionaries meant for human beings), creating mappings between tense aspect modality labels, vibhaktis etc. Work will also be needed in lexical semantics: analysing relationships among different senses of word and how they arise, and in discovering universals about them. Also required is a comparative study of language constructions and phenomena, and how they map onto each other across languages.

Computer scientists will have to work on the basic LA software and suitable interfaces. The task of the basic software will be to analyse and generate text. It will include a morph package (morphological analyser and generator), local word grouper and splitter, routines for accessing dictionaries, resolving lexical category and word sense clashes etc. There is need for coming up with an appropriate architecture for the system. Suitable interfaces are needed for two categories of people: for the users of LA for accessing a document in another language, and for the linguists who wish to use it as a tool for doing comparative analysis of languages or to look at phenomena in other languages.

Finally, there are certain infrastructural needs. First, for LA to be useful, documents need to be available in electronic form. Today, large numbers of documents, including many books, and perhaps all newspapers in Indian languages are composed by the computer. As a result, the text is already available in machine readable form. There is need to create a system for its systematic distribution. Second, there is need to create computer facilities in libraries for example, on which LA software can be run. The community of users can contribute (and get gainful employment) by post-editing the output generated by the LA system, and producing translations for use by others. These can then also see the print media.

## 6. Conclusion and future work

We have described the Paninian framework as adapted to modern Indian languages. It turns out to be elegant as well as natural and efficient. It is elegant because it is able to handle diverse phenomena like karaka assignment, active-passive, and control in a unified manner. It is efficient because the constraint parser which arises very naturally using the framework is extremely fast.

Finally, we talk about the concept of anusaraka. It makes possible the building of translation aids extremely fast, at least for Indian languages. Currently, a 30,000 word anusaarak from Kannada to Hindi has been built and tested in our laboratory.

We gratefully acknowledge the help received from Dr K V Ramakrishnamacharyulu of Rashtriya Sanskrit Vidyapeetham, Tirupati in development of the theory. Several people have worked on the implementation of the core parser: Sivasubramanian, B Srinivas, P V Ravisankar on the earlier version, and Jayvant Anantpur, Vasudev Verma, Amba Kulkarni on the current version. Mr V N Narayana has been working on the Kannada-Hindi anusaraka.

## References

- Bhanumati B 1989 An approach to machine translation among Indian languages, Tech. Rep. TRCS-89-90, Dept. Comput. Sci Eng., Indian Institute of Technology, Kanpur
- Bharati A, Chaitanya V, Sangal R 1990a A computational framework for Indian languages. Course Notes for Intensive Course in NLP, Vol. 1, Tech. Rep. TRCS-90-100, Dept. Comput. Sci. Eng., Indian Institute of Technology, Kanpur
- Bharati A, Chaitanya V, Sangal R 1991a A computational grammar for Indian languages processing. *J. Indian Linguistics* 52(1-4): 91-103
- Bharati A, Chaitanya V, Sangal R 1991b Local word grouping and its relevance to Indian languages. In *Frontiers in knowledge based computing (KBCS90)* (eds) V P Bhatkar, K M Rege (New Delhi: Narosa) pp. 277-296
- Bharati A, Chaitanya V, Sangal R 1994 *Natural language processing: A Paninian perspective* (New Delhi: Prentice Hall of India) (to be published)
- Bharati A, Sangal R 1990 A karaka based approach to parsing of Indian languages. *Proc. Int. Conf. Comput. Languages - COLING 90* (Helsinki: Assoc. Comput. Linguistics)
- Bharati A, Sangal R, Chaitanya V 1990b Natural language processing, complexity theory and logic. In *Foundations of software technology and theoretical computer science 10. Lecture Notes in Computer Science - 472* (eds) K V Nori, C E Veni Madhavan (Berlin: Springer Verlag)
- Kiparsky P 1982 *Some theoretical problems in Panini's grammar* (Poona: Bhandakar Oriental Research Institute)
- Narayana V N 1994 *Anusaraka: A device to overcome the language barrier*, Ph D thesis, Dept. Comput. Sci. Eng., Indian Institute of Technology, Kanpur (submitted)



## **Reducing barriers of communication across Indian languages: An AI and ES approach to mass media**

SHREESH CHAUDHARY

Department of Humanities & Social Sciences, Indian Institute of Technology, Madras 600 036, India

**Abstract.** The present paper identifies some nonlinguistic and linguistic barriers that will have to be overcome by any system for automatic and simultaneous communication of news, commercial advertisements, and other items of information and entertainment by mass media across some Indian languages.

The paper also presents a brief account of some theories for the representation of knowledge of language in a language-independent manner, because such theories can make simultaneous communication of an item across different languages rather easy. But more research is required in this field before the relevant knowledge can be represented in a language-independent manner. Presently automatic and simultaneous communication of an item from any language to many other languages does not seem easy. However, it seems that in a very limited way a beginning can be made in the direction of such communication by human editors aided by tools developed by computer scientists so far.

**Keywords.** Attention-worthiness; barrier; communication; database; knowledge; language; mass media.

### **1. Introduction**

Newspapers, radio and television, the three most widely used media of mass communication, presently take considerable time and human and other resources for preparing language specific editions of the same or similar stories, features and commercials. Efficiency, economy and the reach of these media can increase substantially if tools are made available to communicate simultaneously across many languages.



Rather than prepare a copy, for instance, in English first and then translate the same into different languages through human translators, it would be economical and fast if copies for different languages could be prepared at the same time, without requiring human translators at all. Such a system of mass communication would carry much more information and entertainment to many more people much faster and at cheaper rates, and would also expose many more customers to different mass media, than possible at present because of the constraints of language.

It seems that a system of communication of this kind is likely to encounter difficulties, some linguistic and some nonlinguistic. In other words, a news item, for instance, published or broadcast in one language may not be understood by readers or listeners in other languages either because of differences between the languages known and those available, or because of differences in the reader/listener's assumed knowledge of the background of any news item in different languages.

In the following sections, the present paper is going to analyse the nature of these barriers in such communication. The paper also identifies some issues that can be addressed by artificial intelligence and expert systems (AI and ES) technologies to overcome these barriers and make such communication across languages possible.

Section 2 describes the difficulties of communication in the mass media arising out of nonlinguistic differences among groups of customers. Section 3 describes some difficulties of communication attributable to linguistic differences, and to different conventions of use of each language for mass communication. The final section of the paper indicates issues to be addressed for overcoming these difficulties for mass media in communication across some Indian languages.

## 2. Nonlinguistic barriers

A news item published or broadcast in Tamil may be difficult to understand for non-Tamil knowing readers/listeners, not only because of their ignorance of Tamil, but also because these readers may not know as much about a DMK politician, for instance, as the Tamil news editor may assume.

The demands on the producers of mass media seem to vary from one political segment to another, or from one economic stratum to another, just as they vary from one language community of readers/listeners to another. This appears to be a much bigger problem than may be readily realized. It is normal for a tabloid paper like *The Daily Mirror* of Britain (and India has its own tabloids like *The Blitz*, *The Telegraph* etc. as well), for instance, to headline Iraq's defeat as 'GOTTCHA' when *The Times* of London may headline this item with something like 'Iraq surrenders'. Any AI and ES technology intended to remove the barriers in simultaneous and automatic mass communication across many languages has a greater challenge here.

Because of the varying perceptions of any (news) item, it seems that adapting *The Times* of London, for instance, for readers of *Le Monde* of Paris may be easier than adapting the former for the readers of *The Sun*, or *The Daily Mirror*, or *The News of the World*, even when *The Times* of London and the three tabloids mentioned here are all in English, and *Le Monde* is in French.

The point is that news items or other features and commercials in mass media in different languages do not have only linguistic differences. Differences among them are induced also by political, economic, social and cultural factors, all of which are more varied and unstable than linguistic differences.

Analysis of multiedition newspapers in the same language even in India shows that

news items can have varying degrees of reader-attention-worthiness, and therefore, there can be varying editorial priorities and practices. As examples below indicate, these editorial practices do have linguistic manifestations. But they seem to be more influenced by the political, cultural and other sociological features of the target customer group or groups of each of these mass media. Given in the examples below are some headlines from two English language newspapers published from the same city, i.e. Madras.

**Example 1.** Same city, same language, different newspapers.

(a) February 6, 1992

*Indian Express*

Headline (HL) 1: Cong-I, Kabul faction get one Assembly seat each  
2: Pak not to allow militants to cross Line of Control  
Anchor HL : PM stalls drug price decontrol

*The Hindu*

HL 1: Amrinder, Cong(I) man elected unopposed  
2: Pakistan Govt. to prevent marchers  
Anchor HL : Temple in Tirunelveli ransacked

(b) February 7, 1992

*Indian Express*

HL 1: As Pak steps up anti-India drive...  
Deal firmly with JKLF-march, govt tells para-military forces  
2: Militants determined to cross border  
Anchor HL TN annual plan outlay enhanced to 1,751 cr

*The Hindu*

HL 1: India asks Big Five to counsel Pakistan  
2: Police action in Punjab  
Poll panel rejects State explanation  
Anchor HL Plan outlay for T. Nadu increased

(c) February 9, 1992

*Indian Express*

HL 1: India Prepared to Meet Pak threat: Solanki  
2: PM Convenes Meeting  
Anchor HL 1: Judges, lawyers told not to attack each other  
2: Railways in a hurry to favour MNC

*The Hindu*

HL 1: India Ready to Meet Pak. Threat: Solanki  
2: Busloads of "marchers" on Pak. side of border  
Anchor HL Strains in Cong (I)-AIADMK ties remain

These differences can be found also between editions of the same newspaper published from different cities. For instance, some headlines from two English language newspapers published from two cities in India are reproduced below.

**Example 2.** Same paper, same language, different cities.

(a) *The Hindu* February 7, 1992

*Hyderabad Edition*

- |           |  |
|-----------|--|
| HL        | 1: India asks big five to counsel Pakistan |
|           | 2: Police action in Punjab                 |
|           | Poll Panel Rejects State Explanation       |
| Anchor HL | Move to give Ayacutdars minor tanks        |

*Madras Edition*

- |           |                                   |
|-----------|-----------------------------------|
| HL        | 1: (Same as in Hyderabad Edition) |
| HL        | 2: (Same as in Hyderabad Edition) |
| Anchor HL | Plan outlay for T. Nadu increased |

(b) *India Express* February 9, 1992

*Hyderabad Edition*

- |           |  |
|-----------|--|
| HL        | 1: India fully prepared to meet any threat: Solanki<br>PM convenes joint meet to discuss Kashmir issue |
|           | 2: Sikh ultras held in Delhi, Calcutta<br>7 killed in Haryana train bomb blast                         |
| Anchor HL | Locomotive Purchase<br>Rlys in a hurry to favour MNC   |

*Madras Edition*

- |           |  |
|-----------|--|
| HL        | 1: India Prepared to Meet Pak threat: Solanki<br>2: PM Convenes Meeting                  |
| Anchor HL | 1: Judges, lawyers told not to attack each other<br>2: Railways in a hurry to favour MNC |

Examples 1 and 2 above very clearly show that there are many nonlinguistic differences between the ways in which the same news item is presented in two newspapers of the same language in the same city, or in two editions of the same newspaper in different cities.

These nonlinguistic differences are not just stylistic differences. These differences can be classified broadly into the stylistic, lexical, and perceptual. Stylistic differences, for instance, include those like *Indian Express* (IE) writing 'Pak' without a dot, as in example 1(a). 2 and *The Hindu* (TH) writing 'Pak' with a dot, as in example 1(c). 1 and example 1(c). 2, for instance.

There are many other differences of this kind. For Tamilnadu, IE writes 'TN', as in 1b. anchor. But TH presents it as 'T.Nadu', as in 1b.anchor itself. As King (1987, pp. 5-6) observes, the human mind is such a powerful processing device that these differences will pass as insignificant minor differences in the mind. However, for a mechanical system, which necessarily requires explicit description of knowledge and processing techniques, these stylistic differences might be significant. For instance,

block capitals have been used for the first letters of the two words 'Big Five' in TH, Madras edition, (1b). Human mind can perceive its significance and the logic behind it readily. That, however, may not be so for any mechanical system. But this stylistic feature will continue to remain important for newspapers, though most Indian languages do not present problems of capital and small letters. That, however, only indicates a range of differences of this kind.

Examples 1 and 2 above indicate that media within the same language too, and more so across languages, tend to use different lexical items to indicate the same thing. Whereas IE has used 'not to allow', see (1a), TH has used 'prevent'. In the same headlines, one has used 'militants', whereas for the same people the other has used 'marchers'. In (1c), IE has used 'prepared', TH has used 'ready'. Thus even among the examples 1 and 2 above, there are numerous differences.

Differences in the perception of the same event or news item can be most baffling for any mechanical system. These differences seem to be governed more by the socio-politico-cultural features of the target group of consumers than by any other. For instance, the event of the unanimous election of two members to the Punjab assembly is reported differently in the two newspapers of the same language published from the same city. IE reports it as 'Kabul faction', the faction of the political party to which the elected member belongs. But TH reports it as 'Amrinder', by the name of the elected member. IE reports it as 'Cong-I, Kabul faction'. But TH changes it partly, and reports the item as 'Amrinder, Cong-I'. Differences of anchor headlines between IE and TH of February 9, 1992, as in example 1(c) above, can be explained only by attributing them to their different perceptions by the editors and consumer groups. For IE readers, the alleged irregularities in the purchase of locomotives by the railways seem important. To the readers of TH, the strain in the ties between Cong-I and AIADMK, the erstwhile political allies, may be of greater importance. There are many other differences of this kind.

Such differences can be found in news, features and commercials pertaining to the same item even in radio and television. Different regional stations may present different items in different ways, even when they may be using the same language. Radio news from All India Radio (AIR) whether in Hindi or Tamil or any other language is likely to differ from one regional station to another.

If there are such differences among media within the same language, they are sure to be more so among media across different languages. These nonlinguistic differences of perception arising out of varying **attention worthiness** of different news, features or other items may be the biggest challenge for any system of communication across different languages.

### 3. Linguistic barriers

A more clearly definable problem in the journalistic field pertains to linguistic differences among media. Problems here arise out of differences of script, phonetics and phonology, lexicon, syntax and semantics, and of journalistic conventions of language use.

In other words, a news item or feature published in the press, or broadcast on radio and television in one language may not be legible or comprehensible in another language either because the script or pronunciation used even for shared lexical items

and/or syntactic features may be different, or because the media may be using different lexical and/or syntactic features for communicating a similar piece of information.

### 3.1 *Script and pronunciation*

Hindi and Urdu newspapers offer a good example of the barrier of communication due to different scripts. Ordinary spoken-Hindi and spoken-Urdu have a significant amount of mutual intelligibility. But in the printed form, this intelligibility is totally lost. Hindi is generally printed in the Devanagiri script, and Urdu in the Arabic. These scripts belong to different script families, and have almost no common features in either the symbols or the diacritics used by them. Hindi, and to a considerable extent this seems true of Urdu as well, is intelligible to a larger variety of people when written in a script known to them. The use of Hindi in the Roman script by the armed forces in India until recently can be cited as an example in this context. Given a common script, Urdu and Hindi may appear to be dialects of the same language. But somehow, perhaps due to some accident of history, scripts have got inseparably associated with the popular identity of languages.

Radio and television have a similar problem at the phonological level. Many languages show dialectal differences in the pronunciations of many words which otherwise appear to be similar. The most obvious example of such a variation is the different pronunciations of the name of one of our ancient languages, namely *Sanskrit*. In some parts of India the name of this language is pronounced as *Samskrit*, and in some others as *Samskrut*. There are arguments to justify all these pronunciations and spellings.

Researches during the last three decades in the phonology of different regional varieties of English strongly suggest that the same language might become unintelligible to different groups of speakers because of differences of pronunciation. For detailed discussions of this problem, see Bansal (1969), Chaudhary (1989), and relevant references therein.

### 3.2 *Lexical items*

Among many languages in India there is much similarity of several classes of lexical items, like words for numbers, units of distance, weights and measures, directions, days, dates, months and many words of common historical origin. Yet even these items become unintelligible across language groups because of differences in pronunciation. Publications in the Common Vocabulary Series of the Central Institute of Indian Languages, Mysore, are good illustrations of this phenomenon of the similarity of certain classes of lexical items and their different pronunciations. For instance, we can see Rajaram (1973) for similar lexical items in Hindi and Tamil.

There are formidable barriers in communication across many Indian languages due to differences of lexical items too. In recent years, possibly under populist programmes, these differences have been deliberately emphasized.

For instance, lexical items used in mass media in Hindi and Urdu are, whenever possible, very different from each other even when use of the same or at least very similar items does appear to be possible in these languages. Promoters of the sectarian identity of Urdu seem to have encouraged the greater use of lexical items of Arabic and Persian origin. The media in Hindi seems to have opted for greater use of Sanskrit words. The following examples from news broadcasts by All India Radio (AIR) only indicate this tendency.

**Example 3.**

- (a) satrah sipaahiyon kii mrityu ho gayii  
"seventeen soldiers of death be went"  
Seventeen soliders were killed.
- (b) satrah sipaahii halaak ho gaye  
"seventeen soldiers killed be went"  
Seventeen soldiers were killed.
- (c) satrah sipaahii maare gaye  
"seventeen soldiers killed went"  
Seventeen soldiers were killed.

Example 3(c) is widely attested in nonspecialist use in both Hindi and Urdu speaking communities. But, for certain reasons including those of the cultural perceptions of the consumer groups, mass media, even government-owned media like the AIR, prefer to use 3(a) for Hindi and 3(b) for Urdu.

Some examples of lexical differences for the same item from some Indian language newspapers are given in examples 4 and 5 below.

**Example 4.**

- (a) kendriyo sarkaarer acomko siDDhaanTo
- (b) kendriiy sarkaar kaa achaanak nirNay
- (c) keendriiya prabhutwapu aakasmika nirNayam
- (d) mattiya arasu tiDiir muDivu
- (e) "central government's sudden decision"

**Example 5.**

- (a) panjaab, aasome widhaan sabhaa o lok sabhaa nirwaacana
- (b) panjaab, asam menbhii widhaan sabhaa, lok sabhaa cunaaw
- (c) panjaabu assaamulaloo widhaana sabhaa look sabhaa ennikalu
- (d) panjaab, asaamilum makkaLawai saTTapeerawai terdalhaL
- (e) "Punjab, Assam (in) (also) legislative assembly (and)

House of People election"

State Assembly and Parliament elections in Punjab and Assam also.

The headline in 4(a) is possible in any Bengali language newspaper, 4(b) in Hindi, 4(c) in Telugu, and 4(d) is from the Tamil newspaper *Dinamani*, Madras. Headlines have been given in the same order of languages in example 5 also.

We can see that among the four languages given here that there are varying degrees of lexical similarity, with Tamil perhaps being the most different from the other three. There are one or two items which are different in all these languages. For instance 'decision' in example 4 has different words in all of these languages. In example 5, 'election' has different words in all of these languages.

Similar differences of lexicon have increased in recent years in many Indian languages, and media in all these languages have been under pressure to abandon common words for those perceived as specific to the cultural identities of their readers

and listeners. These differences can be seen clearly in the language specific translation of words and phrases borrowed from English and commonly used in many Indian languages. Some examples of differences in words of this kind have been given in example 6 below.

### Example 6.

Hindi	Tamil	Gloss
Patraalaya	Tapaalnilayam	Post office
Duurdarshan	TolaikaTci	Television
Duurbhash	Tolaipeesi	Telephone
Betaar	Tandi	Telegram
Wimaanpattan	WimaanatLam	Airport
Kendriya sarkaar	Mattiya arasu	Central Govt.

Thus, words of even common historical origin are getting localised, and to that extent, communication across languages even of the same family is becoming difficult.

### 3.3 Syntax

Another difficulty in communication of this kind appears to be there in the differences of syntactic features of various languages and in differences of conventions of use of these features in the media. It is well known that India has several groups of languages belonging to several language families which are significantly different from one another in their typological features.

There are, for instance, Naga, Mizo and some other languages of the North-East. Many of these languages are mutually unintelligible even among themselves. Then there are languages like Kashmiri, Laddakhi and a few others in the North-West which are very different from other Indian languages in their typological features. There are languages like Mundari, Oraon, Kurukh, Santhal, and a few others spoken by the so-called tribal communities in the plateau of Central India which have little in common with even other tribal languages.

Among many languages listed in the eighth schedule to the Indian Constitution, such as Bengali, Hindi, Marathi, Tamil, Telugu etc. however, there is a considerable amount of typological similarity. They are, for instance, mostly of the subject-object-verb (SOV) type, unlike, for instance, an SVO type language like English. Many of these languages also have grammatical features of number, gender, person, case, honorifics and tense, though the exact forms of these features differ significantly.

In Telugu, for instance, case is realized through inflexions, whereas in Hindi it is realized through lexical items. The phenomenon of grammatical gender in Hindi/Urdu is much more pervasive than in other Indian languages. Many other Indian languages do not have this phenomenon of grammatical gender. We have already seen some instances of these differences in examples 4 and 5 above. These differences are significant barriers for any medium of mass communication across these languages.

### 3.4 Journalese

In addition to lexical and syntactic differences among these languages, mass media in each of these languages also have specific conventions of language use, with their own syntactic and semantic features. From examples 4 and 5 above, it seems that some newspapers in Indian languages prefer nominal and nonfinite forms to verbal or finite ones for headlines and the first lead. But, for the second lead, their preferences for finite and nonfinite forms differ, as may be seen in the following examples from Bangla, Hindi and Tamil, in that order in each.

#### Example 7.

- (a) ameerikaar tattaawdhaane aalocnaa
- (b) ameerikaa kii dekhrekx men honii caahiye
- (c) amerikkaa meerpaarwayil naDakka weeNDum
- (d) "America (of) auspices (in) (should be) (conference)"  
(Conference) should be under American auspices

#### Example 8.

- (a) arab raashtro shonge shorashori aalocnaar jonne izraael ekmoT
- (b) arab raashtron se siiDhii baaTciiT ke liye izraael sahmat
- (c) arabu naaDuhaLuDan neeraDi peeccu isreel sammadam
- (d) "Arab nations with direct talk Israel agreement"

Israel Agrees for Direct Talks with Arab Countries

Whereas English language newspapers in India show almost an equal preference for nonfinite and finite forms, simple present tense being their preferred finite form, in Indian languages the situation in this regard does not seem to be so clear. Newspapers in these languages demand at least one of the two headlines, and there usually are two headlines for many items, in the finite form, usually the simple past.

Hence, a system intending to facilitate communication across different languages faces a two-fold problem: first, deciding about which headline has to have a finite form, and then converting the finite or nonfinite form of headline in one language to a finite or nonfinite form in another.

Another problem for mass media in communication across languages is in the different linguistic conventions for journalistic and other references to sources of information. In many Indian languages there are already standard phrases for such journalistic expressions as – reportedly, allegedly, according to informed sources, according to reports reaching here, according to a spokesman, a spokesman/person said etc.

But not all languages have the same conventions of using these phrases in the beginning, the end or the middle of a sentence, as the following examples indicate:

#### Example 9.

- (a) It (Union Government) has left it to the Election  
Commission to decide the dates, a spokesman of the  
Home Ministry said tonight.



- (b) eko sorkaarii probokTaa moTe nirbaacon komiishan  
 "a govt spokesman said election commission"  
 eijonne nirbaaconer Diin ghoshnaa korben  
 "for this election for dates announce will"
- (c) ek sarkaari prawakTaa ne baTaayaa ki cunnaw aayog  
 "a govt spokesman said that election commission"  
 ko iske liye TiThiyaan nirDhaariT karne ko kahaa gayaa hai  
 "to this for dates fixed do to said has been"
- (d) idarkaana tedihaLai teerdal kamishan muDiwu seyyum enru  
 "for this dates election commission decide will make that"  
 arasin peeccaaLar oruwar teriwittaar  
 "of govt spokesman one said"

The example in 9(a) is from the English newspaper *The Hindu*, Madras, and the one in 9(d) is from the Tamil daily newspaper *Dinamani*, Madras. Those in 9(b) and 9(c) are possible in Bangla and Hindi newspapers. These examples indicate that newspapers, and possibly other media too, in many Indian languages do not follow the same conventions for acknowledging their sources of information.

To summarise then, the following are some of the major issues that any system for simultaneous communication across some Indian languages may have to encounter.

#### *Nonlinguistic*

- (a) Varying attention-worthiness of items according to groups of readers/listeners/viewers.
- (b) Varying amounts of complementary information required for different items according to groups of readers/listeners.

#### *Linguistic*

- (a) Different scripts for the same language.
- (b) Different pronunciations of the same item.
- (c) Different lexical items for the same thing.
- (d) Different rules of sentence construction.
- (e) Different conventions of language use in mass media.

Overcoming these difficulties in simultaneous communication across several languages may not be easy in the near future. However, progress in Artificial Intelligence and Expert Systems and in other fields of computer science indicate that some success can be achieved with the proper use of tools developed so far.

#### **4. Knowledge of language**

The solution to the problem of communication across several languages appears to depend upon the description of what constitutes the knowledge of language. It appears that application of this knowledge may presuppose a description and quantification

of this knowledge, possibly in machine representable form. But what kind of information constitutes knowledge of language?

Can we say that the knowledge of words, letters and sounds in them, their meanings and rules of usage, constitutes the knowledge of language? Or, can we say that to know the rules of sentence construction, as given in the grammar book of any language, constitutes the knowledge of language? Or, can we say that this knowledge is of a very abstract kind not specific to any language but is a common property of the human mind which enables nearly all human beings to learn at least one language?

Since the late 1950s, Noam Chomsky has pioneered a new approach to the study of the structure of language and of how the knowledge of language seems to be acquired and organised. Chomsky (1981, p. 4) argues that knowledge of language essentially comprises a knowledge of its rules and representations. It is this body of rules and representations, according to Chomsky (1986, pp. 12–13), which “relates sound and meaning and assigns structural properties to physical events in certain ways, not others”.

Chomsky (1986, pp. 12–13), therefore, believes that “We should think of knowledge of language as a certain state of the mind/brain, a relatively stable element in transitory mental states once it is attained; furthermore, as a state of some distinguishable faculty of the mind – the language faculty – with its specific properties, structure and organisation, one ‘module’ of the mind”. For a detailed description of this view of the knowledge of language, popularly known as ‘Generative Linguistics’ or ‘Generative Grammar’, we may see Chomsky (1981, p. 4, 1986, pp. 12–13) and other references.

This view of language may be very convenient for any system of application as it presupposes only one underlying representation for all languages of the world. It also assumes that representation of knowledge of language is not language-specific. To take a crude example, it is like representing the knowledge of an English word like cat with a picture of a four-legged creature. Such a representation enables easy and instant access to knowledge in a very language-free manner. Given such a representation of knowledge, any system for communication across languages will reduce all texts from any language to this form of knowledge, i.e. a visual form, and then generate the same text through any number of desired languages.

During the last 30 years, many languages have been described within the framework of generative linguistics, and a considerable body of knowledge about natural languages has been built up. But these works are so lacking in some crucial details that they are considered to be of limited use for engineering in natural languages, or for application of this knowledge for any practical purpose.

For instance, the entire phonological component of the generative linguistics is based on a presupposition that there is a discrete unit of sound, called *phoneme*, even in continuous speech such that it is distinct from other such units and can always be isolated. Researches in segmentation of continuous speech, however, have shown that it is extremely difficult, or nearly impossible, to indicate the particular point where one sound ends and the other begins in a normal continuous utterance,

The human mind nevertheless recognises some discrete sound units, irrespective of variations and other physiological factors. But no reliable description seems to be available so far of how exactly this recognition takes place. For want of such a description, studies in the field of speech synthesis and similar fields of engineering and application of knowledge of speech have used a term like ‘diphone’ which includes adjacent parts of two discrete sounds (see Klatt 1987). Consequent upon our inability

to identify the phoneme, studies in speech engineering have also used clusters of sounds upto the maximal unit of a syllable for identifying discrete sound sequences. Works like the ones reported in Yegnanarayana *et al* (1991, pp. 467–475) make use of a large number of sound sequence units, nearly 350, for speech synthesis in Hindi, whereas any language requires typically about 45 sound units (phonemes).

The approach which codifies knowledge of speech through diphones and/or discrete representation of different syllables may be useful from the engineering and application points of view, but appears somewhat *ad hoc* theoretically and also cumbersome in practice. It simply ignores the question of quantifying the knowledge that enables the human mind to recognise the beginning and the end of one sound from those of another, or recognise the presence of other sounds in any other utterance. From an engineering point of view also such an approach does not seem to be efficient as it increases the processing time and cost. For a discussion of engineering issues in this field see Klatt (1987).

In recent years, however, there have been attempts to write rules and representations in machine usable form for speech understanding and natural language processing. Ostensibly in the form of conventional parsing rules, these rules and representations are logically much more constrained, have mathematical precision, and describe many abstract concepts with an accuracy not available earlier. For illustration, we shall refer to two of these attempts here.

One of these, first reported in Schank (1973, pp. 45–6 & 192–4), has to do with the ‘Identification of conceptualizations underlying natural languages’. The work reported here addresses itself to the problem of the ‘representation of meaning in an unambiguous language-free manner’. Schank believes that ‘there exists a conceptual base into which utterances in natural languages are mapped during understanding’. So, for instance, Schank (1973, pp. 45–6 & 192–4) suggests that we can refer to a nominal concept as a “Picture Producer (PP) such that it is the concept of a general thing, a man, a duck, or the Grand Canyon... An action (ACT) is what a nominal can be said to be doing. For a concept to qualify as an action it must be something that an animate nominal can do to some object... the heart of any conceptualization is the relationship between the actor and the action in the event”. Following this approach, Schank can paraphrase the difference between the following pairs of sentences:

#### **Example 10.**

- (a) John ate the steak with the odor.
- (b) John ate the steak with the fork.

#### **Example 11.**

- (a) John shot the girl with a rifle.
- (b) John shot the girl with long hair.

For details, we can see Schank (1973, pp. 45–6 & 192–4). However, this approach to machine understanding of natural languages has its own limitations. It seems to have limited efficiency with sentences which refer to abstract concepts, or to information preceding or following in the discourse. For instance, it may be difficult for this approach to paraphrase the difference in the semantic roles of John in the following sentences:

**Example 12.**

- (a) John is easy to please.
- (b) John is eager to please.

In 12(a) John is easy for someone else, not mentioned in the sentence, to please. In 12(b) John is eager to please someone else, not mentioned in the sentence. All these may require extra-sentential knowledge along with knowledge to paraphrase abstract concepts like "please" in a language-free manner.

There are other kinds of difficulties with the identification of conceptualizations underlying natural languages. These difficulties pertain to, as mentioned by King (1987), the knowledge specific to the culture of any language community. Some of these difficulties also stem from idioms, unfinished sentences, or assumptions and such other deformities of discourse in any natural language. It seems that these deformities account for a greater part of natural language use. Nevertheless, the approach indicated in Schank (1973, pp. 45-6 & 192-4) holds considerable promise.

Goodman & Reddy (1980, pp. 234-46) in a later work of this kind are more specific about a possible model for "Alternative control structures for speech understanding systems". They assume that "control structures are an essential part of any speech recognition system. They are the devices by which passive knowledge about the task and language is transferred into active and effective processes".

Goodman & Reddy (1980, pp. 234-46) assume that control structure may be divided into three areas: knowledge source interaction, knowledge source activation, and knowledge source focussing. The sources of knowledge (KS), according to them, include the characteristics of speech sound (*phonetics*), variability in phonetics (*phonology*), the stress and intonation patterns of speech (*prosody*), the sound patterns of words (*lexicon*), the grammatical structure of language (*syntax*), the meaning of words and sentences (*semantics*), and the context of the conversation (*pragmatics*).

Goodman & Reddy (1980, pp. 234-46) then go on to propose various models of KS interaction, activation and focussing. For a model of natural language understanding in a language-free manner their proposal seems to be very close to the model of universal grammar assumed in generative linguistics in Chomsky (1981, p. 4, 1986, pp. 12-13), and elsewhere. If and when this model of representation of knowledge succeeds, it should be possible to reduce the text into language-free representations and then generate the same text in any number of desired languages.

There, however, are certain areas of natural language understanding where this model may not be very helpful. Most difficulties of this model have to do with inter-componential knowledge. It may not be easy, for instance, to decide if word stress pertains to lexicon or prosody, or whether case inflexions pertain to syntax, lexicon or phonology etc. It may be difficult within this model to say if the phenomenon of stress shift in words like *photograph*, *photography* and *photographic* can be decided by prosody, or by phonology, or by lexicon. This model does not appear to be adequate for understanding of overlapping phenomena of the kinds mentioned here.

Ignoring certain gaps of detail in the generative linguistics point of view of the knowledge of language, it seems logical to say, that a multilingual person creates a module, or a system of knowledge, which includes knowledge of different languages and the knowledge of different parts of these languages, such as phonetics, phonology, morphology, lexicon, syntax, semantics and pragmatics, and their mutual relationships, as subsystems or as sub-subsystems.

In this module all subsystems, such as those of sound, word, phrase, sentence and

discourse structure, are related to one another. Hence, given an utterance of any size, even with distortion, it seems that the module of rules and representations built by the multilingual person is immediately activated, and then the module initiates multilateral processing of the linguistic data most of the time to generate an appropriate verbal or nonverbal response. This module is so efficient that it handles even complex input correctly and quickly.

To be able to recognise and use different languages automatically, like a multilingual human being, a computer or any other potentially intelligent device will require multiple, such as visual, auditory, olfactory, and other sensory and nonsensory channels of data input, a large, though adequately constrained, lexicon, and inferential and evaluative rules of global reach for data processing. In the present state of knowledge and engineering in this field it does not seem possible.

A fair idea of difficulties in such a task can be had from the account by Steiner *et al* (1988) of the progress of the European Community supported project of simultaneous machine translation in eight European languages including English, French, German, Italian, Portuguese and Spanish.

Engineering in the field of neural networks (see Yegnanarayana 1991, pp. 103–29, for instance), however, seems to be a promising development in this area. But it may still be many years before we can justifiably place any hopes on it. To be able to match the computing powers of the human mind, it seems that a machine will not only require multiple channels for receiving data, such as visual, auditory etc., but also multitier simultaneous processing. Such a machine will also require the power of nonlinear, nonsequential and fuzzy-logic computation for processing the data and then for responding to it in an appropriate manner, occasionally on the basis of instincts and intuition like human beings. Clearly, computers of the power required for this purpose are not in sight yet.

So what can be done to facilitate communication across different languages through mass media at the moment? It appears that presently human editors using the available AI and ES technology can to some extent accomplish simultaneous communication of an item across several languages. An outline of such a solution is given in the following paragraphs.

## 5. Towards a solution

### 5.1 *Nonlinguistic barriers*

In the present state of the art, human editors seem indispensable for deciding what item of information or entertainment may be of interest to which group of customers. But once they decide on such items, providing complementary information about these items can be made significantly less difficult by the use of databases for storing, ordering and retrieving data on any given item in any given language.

Editors in the mass media for cross-linguistic communication may create and maintain some kind of a multilingual database on people, places, events, clientele etc. Monolingual databases are already in use by large business and public organisations for storing and retrieving information about their personnel, operations, clients etc. Multilingual databases are also possible now. Databases can be built and maintained in as many languages as the particular media organisation may wish to cater to.

These databases may be created on the basis of the frequency of the occurrence

of an item in a given language. If, for instance, a particular name occurs quite frequently in a language, then the amount of complementary information required for that item in that language is likely to be smaller than the amount of information required for this item in a language where it does not occur so frequently. To take a concrete instance, let us say that the name of a politician called Yadav is mentioned frequently in Hindi news media for clients in Bihar, but does not occur so frequently in the Tamil news media for clients in Tamilnadu. It is then logical that the amount of complementary information about Yadav in Tamil should be larger than that for him in Hindi. This can be a good working principle for editors to create their databases on a large number of items, i.e. people, places, events etc., and to use them accordingly.

Therefore, once the human editor decides to take up a particular news item in a particular language, all that he may need to do is to retrieve the information stored in the database about that item and put it in the script which can be published or broadcast as appropriate. With the phenomenally large storage capacity of modern computers, it may not be a problem to build databases of different kinds and sizes on a large number of items in several languages. This, however, may only be a tool which can enhance the efficiency and speed of the human editor in the goal of simultaneous communication of information or entertainment through several languages.

## 5.2 *Linguistic barriers – difficulty of differing scripts*

There are languages like Hindi and Urdu where communication is impeded largely due to different scripts being used for writing. In the spoken mode, these languages are mutually intelligible to their speakers. Therefore, with the machine transcription of text from the Devanagiri to the Arabic script, or vice versa, mutual intelligibility can be extended between these languages to their printed mode as well. Unfortunately, the arrangement of letters and diacritics in these scripts do not match. Yet, with the state-of-the-art in computer science, it should not be impossible to design an expert system for machine transcription of texts with human participation wherever necessary.

In the other major Indian languages, fortunately, there is a large extent of similarity between the arrangement of letters and diacritics so that it should be relatively easy to obtain such transcriptions among these languages. For instance, the arrangement of letters and diacritics in the scripts of Telugu and Kannada has a great degree of similarity. If, somehow, the text in one language can be transcribed into the other, a great deal of intelligibility can be obtained for the reader of the target language in this way as well. A similar facility can be created through transcription of texts between Hindi and Punjabi, between Bengali and Oriya, between Tamil and Malayalam etc.

Bilingual keyboards are already in use. All we need is a technology to identify and match letter and diacritics.

## 5.3 *Difficulty of different lexical items & pronunciations*

Sometimes the same lexical item with the same meaning also becomes unintelligible to users of related languages in India because of the different pronunciations of the particular lexical item in these languages. As an example we saw earlier how the name of one of the ancient languages of India is pronounced as 'Sanskrit', 'Samskrit' or 'Samskrut'. There are large number of lexical items, such as numbers, names of days, dates, months, distance, directions etc., which differ in similar ways. Overcoming

these difficulties of differing pronunciations can also make simultaneous communication across several languages considerably easy.

Advances in multimedia technology can be of use here. They are already being used for educational purposes in European countries. In India too, language-specific pronunciations of frequently occurring lexical items can be stored in laser disc-based multimedia systems, and can be used when needed both by the print and the voice media. This approach to lexical differences among related languages can greatly reduce the barriers of simultaneous communication across several languages.

For instance, language-specific lexical terms for the word 'five' can be stored in the multimedia memory, both in print and in voice. This can be retrieved by the human editor for print or voice media according to need. Building such databases no longer appears impossible. Similar things are already being provided in terms of value-added services by personal computer manufacturers. As we have seen earlier, lack of mutual intelligibility at least among major Indian languages is not so much due to differences of syntax as due to differences of lexical items. A database of lexical items, or an electronic dictionary, so to say, of languages, with cross language matching and referencing of this kind, especially in multimedia as suggested here, can greatly enhance the mutual intelligibility among these languages, and also facilitate simultaneous communication across them.

#### 5.4 *Differences of syntax*

We have seen earlier that the major Indian languages do not have significant syntactic differences. They are all of the subject-object-verb type, the phrases in the sentences of these languages are accordingly arranged, and almost a perfect match can be obtained between sentences in these languages.

The major difference, however, lies in the way case marks are shown in the phrases of these languages. In Hindi, for instance, case marks are generally shown through discrete lexical items, whereas in many other major India languages case marks are shown through rule-governed inflections. This is a major problem and a hurdle between simultaneous and automatic communication of an item across these languages.

If the machines can identify the boundaries of a phrase/word, then substituting that phrase/word in one language with the corresponding word/phrase from the target languages(s) cannot be very difficult, especially with the database of lexical items mentioned above, and human editors participating in it. Bharati *et al* (1991, pp. 277-93) and Rao & Yegnanarayana (1991) have also shown that computer technology can be used for identifying word boundaries in Hindi. Once such recognition takes place, the human editor will only have to find the corresponding item from the target languages(s) with help of the lexical database, and make appropriate substitutions. As King (1987, pp. 5-6) has shown, something similar is already being done by the Canadian Radio Organisation for making bilingual weather forecasts. In this manner, computerised databases and laser disc-based multimedia can become effective tools in the hands of the human editor in his objective of achieving simultaneous communication across several languages.

#### 5.5 *Difficulty of different journalistic conventions*

Evolution of journalistic conventions in different languages has been influenced by many factors. Of these, two very important factors are the cultural perception of the

community using that language, and the constraints of space and time on the human editors of the media in that language at any given time.

It seems entirely a matter of cultural perception why for instance, *Union Government* is called *Central Government* (Kendriya Sarkar) in Hindi, whereas it is *Middle Government* (Matthiyaarasu) in Tamil. These terms can be stored and retrieved in a database in the manner stated above for each of the different languages used by any mass media organisation. A human editor will still have to participate and conduct the search, but the electronic dictionary mentioned above can greatly facilitate this search and make cross-linguistic communication less difficult and less expensive.

What, however, cannot be accomplished so easily is the conversion of journalistic items like sources of information from one language to another. For instance, it cannot be decided *a priori* which of the following should be used.

### Example 13.

- (a) a spokesman of the government said
- (b) according to a spokesman of the government

The choice of 13(a) or (b) will depend on the context and the space/time available with the editor. In a business where every inch of space and every second of time counts, it may be of vital importance to the editor to decide which of the competing terms will be appropriate in the given context. But once the human editor decides to use a term of a certain length and kind, then he can look for matching terms in the database of the particular language with him, and use the most appropriate term. Many other decisions with regard to journalistic conventions will also have to be made in a similar manner.

## 6. Conclusions

Automatic and unrestricted simultaneous communication across languages through the mass media does not seem possible presently in view of some of the more formidable barriers mentioned in this paper. With the knowledge available in computers and in the AI and ES fields so far, certain kinds of communication across some Indian languages seem possible even now. Many (news) items, features and commercials involving only names, numbers, dates, days, months, and such other items can, even in the present state of knowledge, be communicated across many Indian languages through the mass media. In other words, the mass media can even now communicate weather forecasts, election results, market rates, company reports, railway, airline and radio and television timetables etc., across several Indian languages simultaneously. Copies of many commercial advertisements for different media can also be prepared in many languages simultaneously. As King (1987, pp. 5-6) reports, with progressively diminishing human participation, Canada has been having simultaneous weather forecasts in French and English from the late nineteen seventies.

I am grateful for the help I have received from Prof B Yegnanarayana and from an anonymous reviewer in the preparation of this article. However, I alone am responsible for mistakes, if any, here.



## References

- Bansal R K 1969 *Intelligibility of Indian English*. Monograph of the Central Institute of English & Foreign Languages, Hyderabad
- Bharati A, Chaitanya V, R Sangal 1991 Local word grouping and its relevance to Indian languages. In *Frontiers in knowledge based computing* (eds) V P Bhatkar, K M Rege (New Delhi: Narosa)
- Chaudhary S C 1989 *Some aspects of the phonology of Indian English* (Ranchi: Jayswal)
- Chomsky N 1981 *Lectures on Government and binding* (Dordrecht: Foris)
- Chomsky N 1986 *Knowledge of language* (New York: Praeger)
- Goodman G, Reddy R 1980 Alternative control structures for speech understanding systems. In *Trends in speech understanding systems* (ed.) W A Lea (New York: Prentice Hall)
- King M 1987 *A tutorial on machine translation* (Geneve, Suisse: ISSCO)
- Klatt D H 1987 Review of text to speech conversion for English. *J. Acoustic Soc. Am.* 82: 737-793
- Rajaram S 1973 *Hindi-Tamil common vocabulary: CIIL Common Vocabulary Series - 3* (Mysore: Central Institute of Indian Languages)
- Rao G V R, Yegnanarayana B 1991 Word boundary hypothesization in Hindi speech. *Comput. Speech. Language* 5: 379-392
- Schank R C 1973 Identification of conceptualizations underlying natural languages. In *Computer models of thought and language* (eds) R C Schank, K M Colby (San Francisco: W H Freeman)
- Steiner E H, Schmidt P, Zelinsky-Wibbelt C 1988 *From syntax to semantics: Insights from machine translation* (London: Pinter)
- Yegnanarayana B 1991 Neural Networks. In *Artificial intelligence & expert system technologies in the Indian context* (ed.) V V S Sarma (New Delhi: Tata-McGraw Hill) vol. 2
- Yegnanarayana B, Murthy H A, Sundar R, Ramachandran V R, Madhukumar A S, Alwar N, Rajendran S 1991 Development of text to speech system for Indian languages. In *Frontiers in knowledge based computing* (eds) V P Bhatkar, K M Rege (New Delhi: Narosa)

## Significance of knowledge sources for a text-to-speech system for Indian languages

B YEGNANARAYANA<sup>1</sup>, S RAJENDRAN, V R RAMACHANDRAN  
and A S MADHUKUMAR

Department of Computer Science and Engineering, Indian Institute of  
Technology, Madras 600 036, India

<sup>1</sup>E-mail: yegna@iitm.ernet.in

**Abstract.** This paper discusses the significance of segmental and prosodic knowledge sources for developing a text-to-speech system for Indian languages. Acoustic parameters such as linear prediction coefficients, formants, pitch and gain are prestored for the basic speech sound units corresponding to the orthographic characters of Hindi. The parameters are concatenated based on the input text. These parameters are modified by stored knowledge sources corresponding to coarticulation, duration and intonation. The coarticulation rules specify the pattern of joining the basic units. The duration rules modify the inherent duration of the basic units based on the linguistic context in which the units occur. The intonation rules specify the overall pitch contour for the utterance (declination or rising contour), fall-rise patterns, resetting phenomena and inherent fundamental frequency of vowels. Appropriate pauses between syntactic units are specified to enhance intelligibility and naturalness.

**Keywords.** Text-to-speech system; prosodic features; coarticulation; intonation; formants; content word; function word.

### 1. Introduction

The function of a text-to-speech system is to convert a symbolic input (text) to an output speech waveform. To produce speech from a given text, human beings use several knowledge sources such as phonetics, phonology, morphology, syntax, semantics and pragmatics. It is necessary to incorporate these knowledge sources in a suitable form for a text-to-speech system to accomplish the same task. Mere concatenation of the signals corresponding to the basic units of speech does not produce intelligible and natural-sounding speech. The rules governing various knowledge sources are essential. Other than the production of isolated utterances of basic units, most of the knowledge sources are acquired by human beings without explicit training or learning. Moreover, these knowledge sources by themselves do not make up speech. Hence these knowledge source can be viewed as metalevel

knowledge. This paper addresses some issues related to the role of various knowledge sources in the development of a text-to-speech system for Indian languages.

Speech is the primary method for communication between human beings. Of the many varieties of life sharing our world, only human beings have developed the vocal means of coding to convey information beyond a rudimentary level. Through the development of a system for speech communication between man and machine, we constitute a whole new range of communication services to extend man's capabilities, serve his social needs, and increase his productivity. As computers become increasingly popular in nearly all segments of society, it is quite natural to consider a natural mode as medium of communication. Speech is obviously the most useful medium of communication between computers and its human users. The other possible method for representing natural communication is in text mode which can be considered as a string of conventional symbols (Allen 1985). Text is often considered a more durable medium of communication and is preserved more reliably. Hence it is widely used for both input and output of computers. But text requires specialized equipment as well as typing and reading skills which many potential users may not possess. On the other hand, speech is the most widely used communication medium between humans and requires no special training. Due to these advantages, there is a growing trend towards the development of speech systems over the past three decades.

Most of the problems that computers currently solve use programs where the steps of solution are defined explicitly. The conventional programs are rigid structurally, their actions are predictable in advance and they cannot handle problems that their programmers did not foresee. But as human beings, we are able to handle and frequently solve problems for which algorithms do not exist and which are characterized by ill structure, ambiguity, incomplete problem understanding, uncertainty and formidable complexity. The ability of human beings to solve such problems is almost taken for granted and is not fully understood. Apart from the unique human ability of common sense reasoning we use various other tools for problem solving such as logic, heuristic search and the extensive use of domain knowledge. To perform natural tasks using computers, one has to program them to exhibit similar problem-solving capabilities, or perhaps in some cases even to surpass human beings. Acquisition and incorporation of domain knowledge which includes formal and empirical components, play a key role in this experiment and hence this approach can be called a knowledge-based approach.

Vision and speech are two primary senses of human beings. Man learns about his environment largely through his eyes and communication is done mainly through the voice. Both the human visual system and the speech production mechanism have their limitations and peculiarities. In order to incorporate the features of these primary senses into machines, we have to formulate a set of rules which consider the possibilities and limitations associated with the task, convert them into a sequence of representable form and incorporate them into a machine in some systematic fashion. Acquisition of various knowledge sources from continuous speech and incorporation of the knowledge in a text-to-speech system demonstrate some aspects of knowledge-based systems related to man-machine communication by voice.

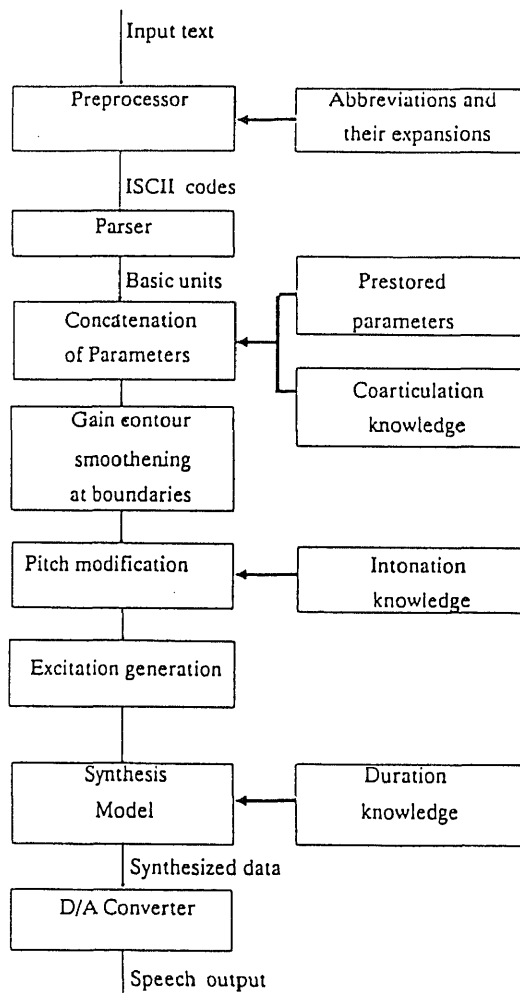
This paper is organized as follows: § 2 discusses the design issues in the development of our text-to-speech system. The role of various knowledge sources in a text-to-speech system for Hindi is discussed in § 3. Section 4 discusses the improvement in the quality of synthetic speech by the addition of these knowledge sources.

## 2. A text-to-speech system for Hindi

We are developing a text-to-speech system for Indian languages based on a parameter concatenation model (Yegnanarayana *et al* 1990, pp. 467–76). As the speech is modelled using parameters, the voice characteristics can be manipulated and thus prosodic features can be incorporated by changing these parameters. The parametric representation is highly flexible and needs much less storage compared to the waveform concatenation model.

The design of our text-to-speech system is modular. It enables us to make changes in any of the modules independently. These modules can be integrated to the rest of the system. Each of the modules added to the system was developed in parallel. The various modules available in our system include the knowledge sources related to coarticulation phenomena (Ramachandran & Yegnanarayana 1992), duration (Rajesh Kumar 1990) and intonation (Madhukumar *et al* 1993).

Figure 1 shows the block diagram of our text-to-speech system. The input to the system is Hindi text stored in the form of ISCII (Indian Script Code for Information



**Figure 1.** Block diagram of a text-to-speech system for Hindi.

Interchange) codes. The preprocessor scans the string of ISCI codes to locate abbreviations, numbers, dates and special symbols and replace them by their expansions in spoken form. Basic units are extracted from the expanded text using a simple parser. For synthesizing speech, parameters of the basic units of input text are concatenated and coarticulation rules that operate across adjacent basic units of speech are applied. The gain contour is smoothed at the boundaries between adjacent basic units. The pitch contour is modified to incorporate the intonation knowledge for the sentence being synthesized. The modified pitch and gain contours are used to generate an excitation signal. The excitation signal and the system parameters are used to generate the speech waveform.

The major issues involved in the design of our text-to-speech system are: (1) choice of basic units, (2) collection of basic units and extraction of parameters, (3) preprocessor and parser and (4) synthesis of speech from the parameters of basic units. In the following sections we discuss each of these in detail.

## 2.1 *Choice of basic units*

The choice of basic units involves a trade-off between the size of memory needed to store all the units and the computation during synthesis. If the size of the unit is large, the number of units in the language increases and hence the computer memory needed to store them is also larger. On the other hand, if the size of the unit is small, then the coarticulation effect among the adjacent units increases, which results in increased computation during synthesis.

For Indian languages, the characters which are generally orthographic representations of speech sounds can be selected as a suitable choice for the basic units. A character in an Indian language is close to a syllable. A character in Hindi represents a speech sound in the form of a consonant (*C*) or a vowel (*V*) or *CV* or *CCV* or *CCCV*. In characters, most of the coarticulation effects (all *CV* and *CC* transitions) are preserved. Also this can be extracted from the text by simple parsing. Due to these reasons characters are chosen as basic units in the present implementation of our text-to-speech system.

The cluster characters can be generated from the constituent *CV* combinations and other consonants. For example, the cluster character /*kya*:/ can be generated by concatenating the consonant /*k*/ and the *CV* combination /*ya*/. This results in the reduction of the number of basic units (from about 5000 to 400) and the storage requirement. Therefore the basic units in our text-to-speech system are: (1) isolated consonants (*C*); (2) isolated vowels (*V*) and (3) the consonant-vowel combinations (*CV*).

## 2.2 *Collection of basic units and extraction of parameters*

The basic units were extracted from the carrier words in isolation. The carrier words selected are meaningless words to avoid the undesirable prosodic bias introduced subconsciously by the speaker. Also it allows us to quickly form a suitable carrier word to make the extraction of the basic units easier. The required basic unit is placed in the word medial position followed by a stop consonant with some exceptions (Rajesh Kumar 1990).

To synthesize speech from a given text, our text-to-speech system uses the following speech parameters: 1) linear predictive coefficients (LPC), 2) formants, 3) pitch and 4) gain. LPC and formants represent the vocal tract system and pitch and gain parameters

correspond to the source information. In the following paragraphs, we discuss the extraction of these parameters briefly.

Our text-to-speech system is based on the linear prediction method. A set of 14 LPC parameters are used to model the vocal tract system. These are computed using the autocorrelation method (Makhoul 1975). The coarticulation effect is manifested in the speech wave mainly as a transition pattern of formants (the resonant frequencies of vocal tract) (Ohman 1966). A difficult signal processing problem is encountered in incorporating the coarticulation rules due to the incompatibility of the parameters used for basic units (LPC) and for specification of the rules (formants). In order to solve this problem, all basic units are converted to a representable scheme in which the vowel regions are stored using formants and the consonant regions using LPC. Formants are extracted using the properties of group delay functions (Yegnanarayana *et al* 1991).

We modify the intonation pattern by incorporating intonation knowledge obtained by analysis of continuous speech in Hindi. The voiced/unvoiced decision of the basic units is stored separately and this is sufficient enough rather than the actual values of pitch. Later, based on the intonation knowledge the pitch contour of the utterances in voiced region is modified. In the unvoiced region the pitch value is taken as zero.

We are using different methods for computing gain for consonants and vowels. Gain contour for each consonant frame is determined from the residual obtained by the autocorrelation method (Makhoul 1975). In the vowel region, gain for each segment is computed as the sum of squared values of the signal. In order to solve the problems due to the incompatibility of gain computation, gain of each basic unit is pre-edited and stored. During the synthesis, after concatenating the parameters, gain contour is smoothed by interpolation across the boundary of the adjacent basic units.

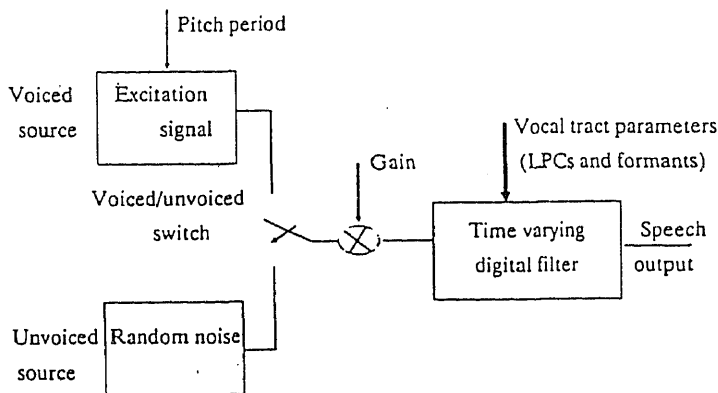
### 2.3 Preprocessor and parser

Preprocessor and parser are two preliminary modules in our text-to-speech system. The input and output of the preprocessor module are in ISCII code itself. The text is preprocessed to locate nonphonetic strings (such as numerals and abbreviations) which are replaced by their spoken form. For example, the abbreviation /*ḍā:*/ (Dr.) expands to its spoken form /*ḍa:kṭar*/ 'Doctor' and the numeral 120.45 expands to /*e:k sau bi:s daṣamlav ca:r pa:ñc*/ 'One hundred and twenty point four five'. It also helps the intonation module to make out if a particular word is a numeral or not. The preprocessed ISCII codes are transferred to the parser.

Due to the phonetic nature of Indian languages, the parser module of our system is simpler than for languages like French and English where letter to sound rules and dictionary look-ups are used (O'Shaughnessy 1984; Allen *et al* 1987). In this module, sequences of ISCII codes are parsed to extract the sequence of basic units. The parser takes care of some language specific issues like word final vowel deletion (for short vowels in Hindi). The end of the sentence is identified by the presence of delimiters (e.g., bar (|), question mark (?) etc.). These sequence of basic units are used for further processing to produce natural-sounding, intelligible synthetic speech.

### 2.4 Synthesis of speech from the parameters of basic units

To synthesize speech from the parameters, we used a linear predictive technique (Atal & Hanauer 1971) in the consonant part and a cascade formant synthesizer (Klatt



**Figure 2.** Block diagram for the basic speech production model. The vocal tract system is represented by LPC and formants. Pitch period and gain are the source parameters. The voiced source of the speech is represented by an excitation signal and the unvoiced source is represented by random noise.

1980) in the vowel part of the basic units. By using this hybrid method for synthesis, it is possible to capture the coarticulation behaviour of speech sounds as a set of transition patterns of formants in the vowel region.

The basic model for speech synthesis is given in figure 2. It consists of an excitation signal and a time-varying filter representing source and system parameters of speech signals, respectively. The excitation signal is periodic for voiced speech signals and a sequence of random numbers for unvoiced sounds. This excitation signal is fed to a time-varying digital filter which models the vocal tract for generating speech. The time-varying filter is represented by either LPC or formants based on the type of basic units.

The choice of excitation signal affects the quality of synthetic speech significantly (Papamichalis 1987). We used Fant's excitation model in our text-to-speech system (Fant 1982). This model is supposed to resemble the actual glottal excitation. Here, the energy is distributed evenly over the entire duration of the excitation. By varying the opening and closing phases of excitation, it is possible to tune the quality of the output speech.

### 3. The role of knowledge sources in a text-to-speech system for Hindi

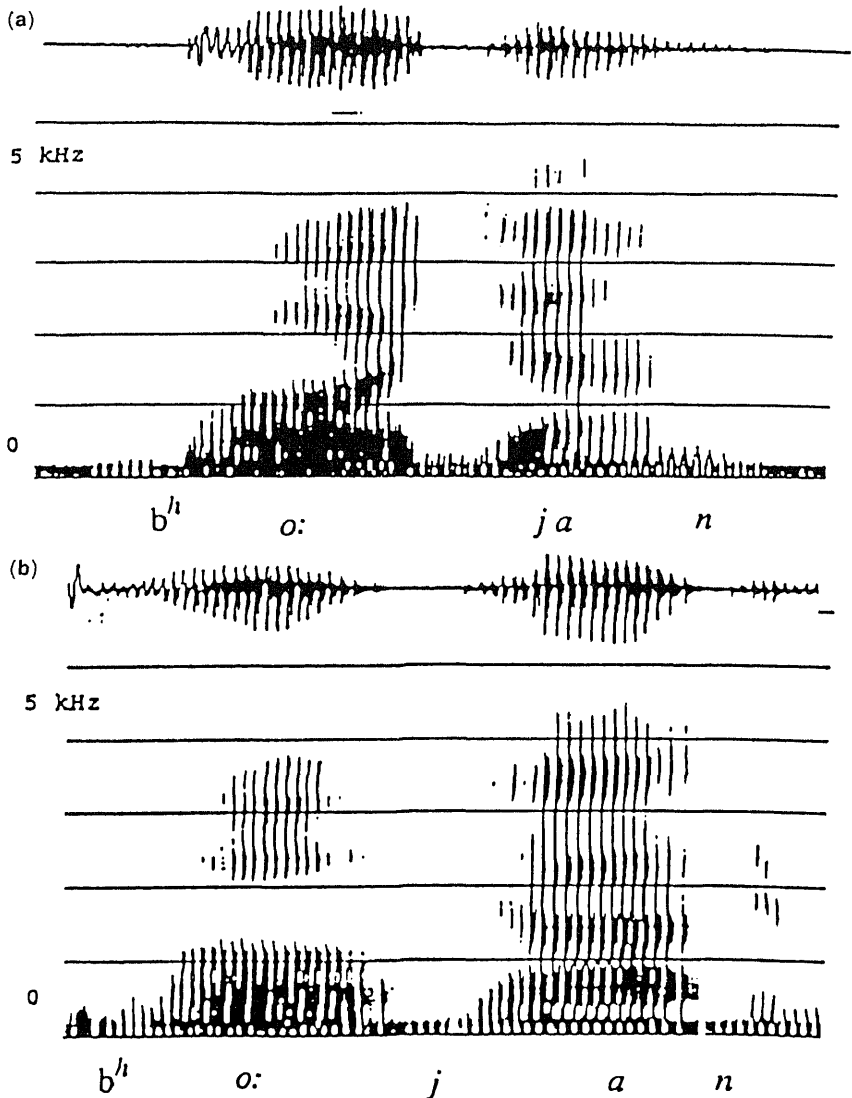
For getting natural and intelligible speech output from a text-to-speech system, incorporation of various knowledge sources at segmental and suprasegmental domains is very important. Segmental-level knowledge sources are concerned with the appropriate representation of speech events by suitable parameters of basic units. Suprasegmental knowledge sources are those whose domains extend beyond a segment.

The important knowledge sources for a text-to-speech system are: (1) coarticulation rules corresponding to changes in acoustic parameters due to the influence of adjacent segments; (2) durational rules corresponding to the inherent duration of segments and their variation due to context; (3) pitch rules corresponding to inherent pitch, pitch variations across words, phrases and sentences, and (4) rules corresponding to

intensity variations across basic units, words and phrases. Rules related to duration, pitch and intensity are classified as prosodic knowledge. In the following sections, we discuss the acquisition and incorporation of knowledge corresponding to coarticulation, duration and intonation.

### 3.1 Coarticulation

The purpose of incorporating coarticulation knowledge in the text-to-speech system is to smooth the transition between adjacent basic units to improve the naturalness and fluency of synthesized speech. Figure 3 shows the waveforms and spectrograms



**Figure 3.** Illustration of the coarticulation effect between speech segments: (a) waveform and spectrogram of the word */b<sup>h</sup> o: ja n/* 'food'; (b) waveform and spectrogram of the characters */b<sup>h</sup> o:/*, */ja/* and */n/* uttered in isolation. The changes in the spectrogram of the vowels of */b<sup>h</sup> o:/* and */ja/* in (a) are due to the context.



of the utterance /b<sup>h</sup>o:jan/ and the isolated characters /b<sup>h</sup>o:/, /ja/ and /n/. The change in formants towards the end of the vowels in /b<sup>h</sup>o:/ and /ja/ are caused by the coarticulation. We study the coarticulation effect in terms of acoustic features like formant transitions, durational changes and intensity variations and formulate rules which predict how these parameters are to be manipulated in various contexts to get the desired coarticulation effect.

**3.1a Nature of coarticulation in Hindi:** The phenomenon of coarticulation involves changes in the articulation and acoustics of a phoneme due to the phonetic context (O'Shaughnessy 1987). The coarticulation phenomenon has an anticipatory and a carry-over component. In the former, features of a phoneme get modified depending upon the following phoneme or phonemes. In Hindi speech, this form of coarticulation is identified to be the following: (1) Change of the features of the vowel in vowel consonant (VC) and vowel to vowel (VV) sequences; (2) the vowel gets nasalized if the preceding or following sound is a nasal, and (3) consonant spectrum is influenced by the following vowel, that is, it may be different for the same consonant before different vowels. Carry-over coarticulation is the one in which features of the phoneme are modified by the preceding phoneme or phonemes. It is observed in Hindi speech as (1) the features of the vowel in a consonant vowel (CV) sequence is affected by the preceding C and (2) the vowel is nasalized to a larger extent if the preceding consonant is nasal.

The acoustic manifestations of the various coarticulation effects are (1) the transition from C to V in a CV unit; (2) the transition from V to C across a VC sequence; (3) the transition from vowel to vowel in a VV sequence; (4) nasalization of the vowel; (5) variation in the duration of the same unit in different contexts, and (6) changes in the spectral features of a consonant in different contexts, namely the CV, VC and CC contexts. Of these, the first three transitions are characterized by formant transition patterns and gain variation across the transition whereas the nasalization is characterized by the presence of antiformants in the vowel spectrum. The CV transition is embedded (or preserved) in the basic unit representation and hence we do not need any contextual rules to make these changes. Our interests are primarily in VC, VV and CC transitions and nasalization of vowels. These are not represented or preserved in basic units since the context which decides the changes is not in the same unit. Hence it is necessary to develop contextual rules to impart coarticulation across unit boundaries and also for nasalization of vowels.

In the study of the coarticulation between two phonemes, we do not consider the influences of segments beyond the immediately neighbouring units. This greatly simplifies the study by reducing the number of cases to be considered. Coarticulation effects due to the phonemes which are not immediately adjacent are less important in fluent speech production though they reduce the speaker's effort in articulation (O'Shaughnessy 1987).

**3.1b Acquisition and formulation of coarticulation knowledge:** The issues involved in the acquisition and formulation of the coarticulation knowledge, from a point of view of incorporating into a text-to-speech system are (1) identification of the domain of coarticulation; (2) classification of the coarticulation patterns, and (3) formulation of the coarticulation patterns.

The domain of coarticulation is the transition between basic units of speech. The transitions between the basic units can be only one of the sequences of vowel-to-vowel (VV), vowel-to-consonant (VC) and consonant-to-consonant (CC).

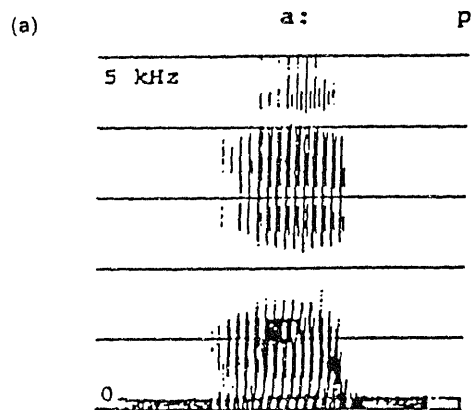
Considering all combinations of two basic units, the total number of junctions possible is about a few hundred. We classify these into a small number of basic transition patterns on the basis of the similarities in the transition patterns. The nature of the transition pattern in a *VC* depends on the articulatory features of the consonant (Ohman 1966). The *VC* transitions are grouped into six distinct classes, each of which corresponds to a place of articulation of the consonant. Each *VC* class is further divided into subgroups based on the manner of articulation of the consonant. Each *VC* subgroup contains formant transition patterns for five different vowels of Hindi. The vowel to vowel formant transition patterns in Hindi are few in number compared to the *VC* transitions. Consonant to consonant transitions (*CC*) are part of cluster characters. The coarticulation across *CC* transitions is more complicated. Few common effects associated with *CC* transitions in Hindi are (1) release of word final cluster, (2) shortening of geminated clusters (geminated cluster is the one in which both consonants are the same), and (3) lengthening of the consonant before a semivowel.

For each *VC* subgroup, rules for transition patterns are formulated for the five different vowels as the percentage deviation of formant frequencies from the steady values and the transition duration. The parameters required for the rule specification are obtained from the analysis of natural speech data. Figure 4 shows the spectrogram of a *VC* transition /a:p/ and also the rules formulated for all bilabial *VC* transitions.

Vowel-to-vowel (*VV*) transitions are characterized by gradual transition of formants from one vowel to the following one. This allows us to take care of all vowel to vowel transitions by a single rule which does the interpolation of vowel formants. Two nasalization rules – one for *CV* and other for *VC* transitions – simulate the nasalization effect. The nasals and nasalized vowels are characterized by the presence of antiresonances in the frequency spectrum. The antiresonance is simulated using a pole-zero pair below the first formant of the vowel. The consonant cluster rules manipulate the duration and the release part of the clusters but do not modify the spectral information of the constituent consonants.

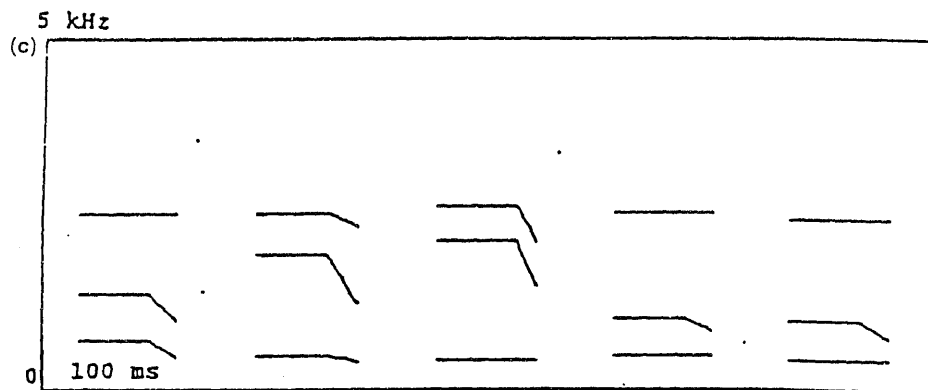
**3.1c Incorporation of coarticulation knowledge:** Incorporation of the coarticulation into a text-to-speech system involves the application of the rules formulated to modify the default parameters of the basic units of the input text before synthesis, by making use of the phonetic context of the basic units in the text. The issues to be addressed in this are those of using the appropriate representation and activation methods. Since coarticulation modifies most of the parameters of the basic unit, it is very convenient to merge the knowledge activation with the synthesis process. The block diagram of the knowledge activation and synthesis scheme is shown in figure 5.

The text is processed sentence by sentence. The prestored representations of all basic units in the sentence are loaded into a buffer (TEMPCVTABLE) so that the parameter modifications by the subsequent processing blocks can be done on this. The analysis of input text is done to decide the phonetic nature of the basic units in the text. This information forms the context for the activation of coarticulation rules. The cluster consonant rule activation process scans the text analysis information and whenever the context for some rule is matched, the rule is activated to modify the default parameters of the concerned basic unit in the TEMPCVTABLE. Next the computation of pitch and gain contours are done using the information in the TEMPCVTABLE and stored in buffers so that the gain and intonation rules can be activated on them before the process of synthesis. The final step is the synthesis along with the activation of transition rules. The *VC* and *VV* rules and the nasalization



(b)

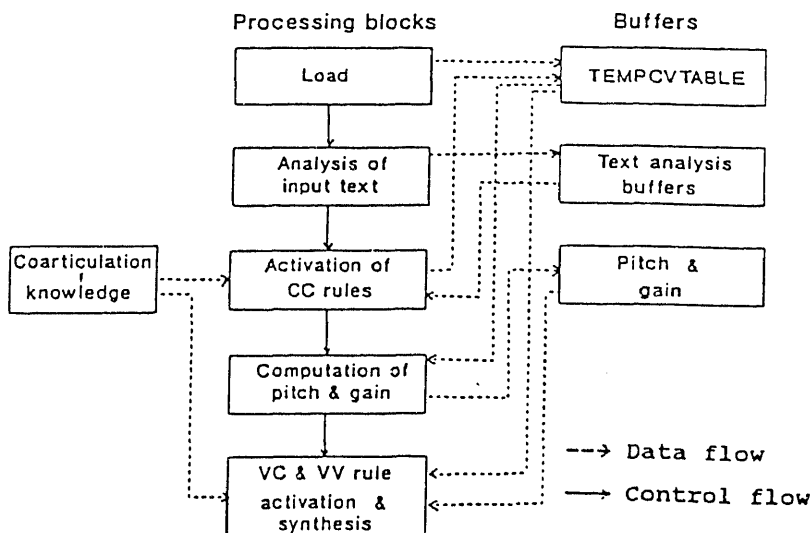
Vowel	a	e	i	o	u
% change in F1	-40	-20	0	0	0
F2	-30	-38	-32	-20	-26
F3	0	-8	-20	0	0
Trans Dur (in frames)	4	4	3	4	4



**Figure 4.** (a) Spectrogram of the VC sequence /a:p/. (b) Table of VC formant transitions from different vowels to the bilabial stops. The transition is specified as the percentage changes in steady formant values of the vowel and the transition duration in frames of 6.4 ms. (c) Formant transition patterns generated from the above table. The first one of these corresponds to the spectrogram in (a).

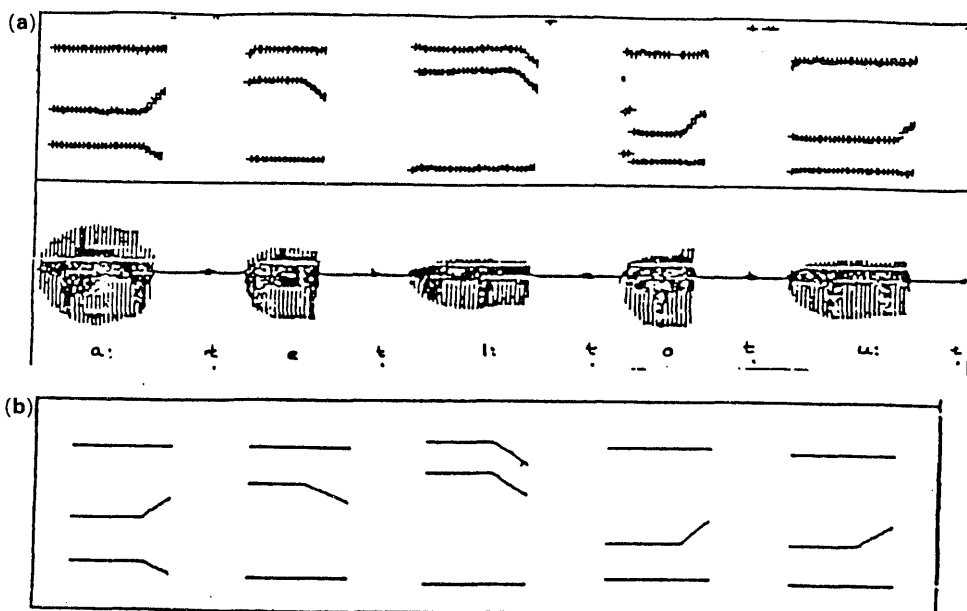
rules are activated in this step. These rules are stored in a table and retrieved and activated as we synthesize the basic units in the TEMPCVTABLE one by one.

The coarticulation rules incorporated into the system are tested to (1) verify their correct activation and (2) evaluate their perceptual significance. This was done by



**Figure 5.** Incorporation of coarticulation knowledge in a text-to-speech system for Hindi.

analytical and perceptual means. Analytical testing of the rules is done by simulating the context in which the rules would be activated and then analysing the synthesized speech by signal processing means. The *VC* and *VV* transition rules and the *CC* or cluster consonant rules are tested. Figure 6 shows formant transition patterns of some



**Figure 6.** Formant transition patterns of *VC* sequences involving /t/. (a) Formant contours extracted from synthesized speech. (b) Formant transition patterns specified by rules.

synthesized VC transitions and the actual transitions specified by the rules. In perceptual testing, words synthesized with and without the transitions incorporated, were heard in pairs. The rules make significant improvement in the perceptual quality. The perceptual evaluation is done at higher levels like in sentences and paragraphs. Selected sentences and paragraphs are synthesized both with and without the coarticulation rules. The speech synthesized with the application of the rules showed significant improvements in naturalness.

### 3.2 Duration

Segmental duration is dependent both on inherent properties of the input unit concerned and on a large number of phonetic and structural constraints imposed contextually (Klatt 1976). The durations of various speech units vary considerably due to factors such as speaker (male vs female), speaking style (reading vs conversational) and speaking rate. Duration of speech units also depends upon the psychological state (fear, anger, sorrow etc.) of the speaker. When a person speaks more slowly than normal, pauses account for more durational increase than the speech units. At faster rates all normal durations of speech units shorten by a certain amount. The interplay of all these factors in natural continuous speech makes duration an extremely difficult feature to study.

**3.2a Nature of duration of speech sounds in Hindi:** For the sake of simplicity, durational effects can be roughly categorized as the effects due to position (POS), syllable boundary (SYL), prepausal lengthening (PPL), post-vocalic consonant (PVC), place of articulation (POA), change in cluster environment (CCL), semantic novelty (NOV), and polysyllabic shortening (PSS). Broadly speaking, the durational effects adjust the durations of basic units depending upon their (i) position and (ii) context in the given text. POS, SYB and PPL modify durations of basic units depending upon their position in the text. Among the durational effects that modify durations of basic units depending upon context, PVC, CCL and POA handle coarticulation at the phonetic boundaries. NOV and PSS cover other contextual phenomena, which are less frequent. In the following sections we discuss each of these effects in detail. These discussions pertain to a limited number of words or nonsense syllables in controlled reading situations by a single speaker. To the extent that the speaker is consistent, effects due to speaking rate are limited. From these studies, rules have been formulated to modify the duration of basic units in the given text.

(i) *Positional effect (POS)* – A character is more lengthened in a word final position than in a word beginning position, which in turn is longer than in a word medial position. The POS results were obtained after analysing durations of about 50 basic units. We proceeded as follows: For each basic unit we formed three nonsense words. Each word contained two or three characters. The first word contained the basic unit (under consideration) in word medial position. The second and the third words contained the basic unit in the word beginning and the word final positions respectively. In these three words, the basic unit is followed by a speech sound of the same category (such as voiced aspirated or voiced unaspirated). This was done to nullify other effects, PVC in particular, so that POS could be studied in isolation. The durations of the basic unit in all three words were measured. The percentage increase of the duration of the basic unit in the word beginning (or word final) position over that

in the word medial position is taken as the value  $\alpha$  for that basic unit for word beginning (or word final) lengthening effect.

(ii) *Syllable boundary effect (SYB)* – The duration of the basic unit appearing just before a syllable boundary is increased. The SYB result was obtained after analysing durations of 24 basic units. We proceeded as follows: For each basic unit, we formed two nonsense words. Each word contained three characters. The first word contained the basic unit in word medial position. The second word contained the basic unit in word initial position with a syllable boundary following it. In these two words, the basic unit is followed by a speech sound of the same category (such as voiced aspirated or voiced unaspirated). The durations of the basic unit in the two words are measured. The percentage increase of the duration of the basic unit in the second word over the first is the combined effect of POS and SYB effects. We remove the POS effect from this in order to obtain the value of  $\alpha$  ( $\alpha$  indicates the percentage by which the 'normal' duration of the unit concerned is modified) for that unit for the SYB effect (it is in this way that we figured out that our rules ought to combine multiplicatively rather than additively).

(iii) *Prepausal lengthening effect (PPL)* – The duration of the character appearing before a pause is increased. PPL effect can be attributed to slowing down of speech in anticipation of a pause, aiding perceptual cues to syntactic boundaries. There is an increase in the duration of either the final or penultimate character of a word just before a pause. If the final character has a vowel, then the increase is only in the final character. Otherwise the increase is in the penultimate character.

The PPL results were obtained after analysing durations of about 20 basic units. This study was performed on continuous speech data and hence the basic units were embedded in meaningful words. The test set consisted of about 25 continuous sentences spoken with natural intonation and rhythm, so that the PPL effects are clearly observed. The duration of each basic unit before a pause (due to either a phrase boundary, or a breath group, or a sentence ending) was measured. The duration of the basic unit, when followed by an unaspirated and an unvoiced stop, is also measured. The percentage increase in the duration of the basic unit in the former over the latter case is the combination of POS and PPL effects. We remove the effect of POS to obtain the value for that basic unit for the PPL effect.

(iv) *Post-vocalic consonant effect (PVC)* – This effect states that the duration of a vowel changes depending upon the type of consonant following it. Voicing, aspiration, sonority and nasality of the PVC, all these affect the duration of the preceding vowel. The PVC results were obtained after analysing the durations of 20 basic units in different contexts. The basic unit could be either a *CV* combination or an isolated vowel. This effect was examined for the vowels /a:/, /i/ and /u/. It was later verified for the remaining vowels. For each basic unit we performed two experiments: (i) PVC effect due to a stop consonant, and (ii) PVC effect due to a nonstop consonant. These two cases are explained.

(a) PVC effect due to a stop consonant: Four nonsense words are formed with the basic unit in the word medial position. The basic unit is followed by four different cases of PVC. They are (1) unvoiced and unaspirated stops; (2) voiced and unaspirated stops; (3) unvoiced and aspirated stops; (4) voiced and aspirated stops.

The percentage increase of the duration of the vocalic portion of the basic unit in (2), (3) and (4) over (1) gives the value for each unit for each category of the PVC.

- (b) *PVC effect due to a nonstop consonant*: Five nonsense words are formed with the basic unit in the word medial position. The basic unit is followed by five different cases of PVC. They are (1) unvoiced and unaspirated stops; (2) trill; (3) fricatives; (4) nasals; (5) semivowels. The percentage increase of the duration of the vocalic portion of the basic unit in (2), (3), (4) and (5) over (1) gives the value of  $\alpha$  for each unit for each category of the PVC. The PVC results were verified in the case of continuous speech for some basic units.

- (v) *Place of articulation effect (POA)* – If two adjacent characters (within as well as across word boundaries) have the same place of articulation, then one or both of the characters are shortened. This is due to relative ease of pronouncing sequences of speech sounds with the same place of articulation.

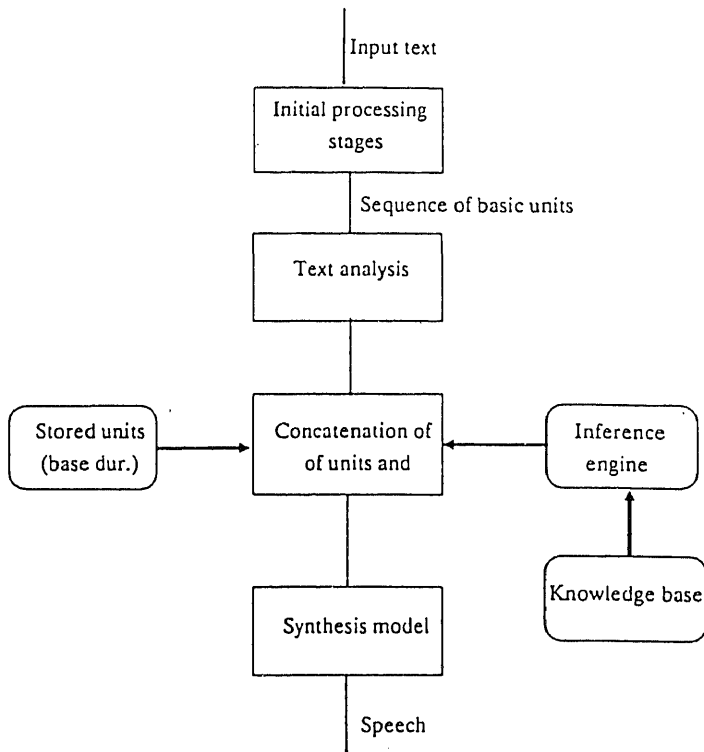
- (vi) *Changes in cluster environments (CCL)* – In the case of cluster characters (CCV or CCCV), the durations of the various constituent basic units change due to the presence of adjacent consonants. They often shorten due to proximity of the POS, and sometimes lengthen due to relative difficulty of pronouncing certain sequences of consonants with conflicting articulatory requirements.

- (vii) *Polysyllabic shortening effect (PSS)* – If the number of characters in a word is greater than three, then the vocalic durations of the various characters are reduced. This effect may relate to communication efficiency: words with many units are easier to identify than short words, which could allow spending less time per unit without risking perceptual mistakes.

**3.2b Incorporation of durational knowledge:** It is possible to vary the duration of a basic unit by varying the number of samples to be synthesized per frame (by default 64 samples are synthesized per frame). For instance, if the number of samples for all frames in a basic unit is doubled, the duration of the basic unit is doubled. The issues involved in incorporation of durational knowledge are (1) analysis of input text; (2) deciding the base duration for each unit; (3) representation and activation of knowledge. Figure 7 places these issues in the overall scheme of our text-of-speech system. In the following sections we discuss each of these issues in detail.

- (i) *Analysis of input text* – The input text is analysed to obtain necessary information to enable the activation of durational knowledge such as (1) type of the basic unit (C or CV), (2) type of consonant in the basic unit, (3) position of the character in the word, (4) number of characters in the word, (5) markers for phrase boundaries and breath groups in the input text, and (6) syllabification within a word.

- (ii) *Deciding the base duration of each unit* – The base duration of each basic unit is its duration in the carrier word where it occurs in the word medial position. Since the same basic unit will be used in all types of context and position wherever it occurs in the input text, it is imperative that the stored basic unit be devoid of the influence of any of the durational effects mentioned earlier. From this point of view some guidelines observed in forming the carrier word for a basic unit, can be explained as follows: (1) The basic unit must be followed by an unaspirated and an unvoiced stop



**Figure 7.** Incorporation of durational knowledge in a text-to-speech system for Hindi.

in order to nullify the PVC effect. (2) Each carrier word must have three characters. This is to nullify the PSS effect. (3) The basic unit and its adjacent characters must not have the same place of articulation. This is to nullify the POA effect. In other words, the base duration of a basic unit is its length in a neutral phonetic context. Depending upon the context in the given text, various durational deviations (using durational rules) are effected. This, in essence, summarizes the durational model used in our present system.

(iii) *Representation and activation of knowledge* – We used the production system approach (Rich 1983) to represent the durational knowledge in our text-to-speech system. Each rule in the knowledge base is an independent fragment of knowledge and does not rely on the correctness of other rules. This facilitates successive updating since the rules are independent of each other, and the order of declaration of rules is not important. For most artificial intelligence application domains, where the knowledge is not systematically formulated (as in our problem), the production system formalism offers a natural way of encoding the knowledge. Besides the production system rules provide an easy way of explaining the intermediate decisions taken.

Since we have represented the durational knowledge as rules, the activation of knowledge is achieved by means of a rule-based inference engine (or a rule interpreter). Depending upon the context of each basic unit in the given text, various rules may be applied. In the modelling duration, it is to be decided whether rules for lengthening



or shortening should be expressed absolutely or as percentages and whether the rules should combine by addition or multiplication. The rules combine multiplicatively if more than one rule fires for the same unit. Thus the order in which the rules combine does not matter. After application of all durational rules, the base duration of each basic unit is modified to obtain its duration to be used during synthesis. The inference engine is of the forward chaining type or is data driven. It is applied for each basic unit in the given text. After all the rules are applied, speech is synthesized using the modified durations of the basic units. The quality of the synthetic speech is improved significantly.

### 3.3 Intonation

Intonation pattern is defined as the variation of the fundamental frequency ( $F_0$ ) with time. An utterance may convey a different meaning due to the changes in intonation even if it is composed of the same segmental phonemes. Intonation helps to group words into syntactic blocks for semantic interpretation of utterances. As in many other languages, intonation patterns in Hindi show some regular features.  $F_0$  contour of declarative sentences decline gradually with time and for interrogative sentences  $F_0$  contour rises towards the end. This backdrop declination or rising is characterized by local falls and rises.  $F_0$  contour gets modified across major syntactic boundaries which is called resetting of  $F_0$  contour. The resetting is used as a marker for phrase boundaries and it is accompanied by a pause. The intonation pattern of an utterance is also affected by segmental factors of constituent units. In the following sections, we discuss the issues in acquisition of intonation knowledge from continuous speech and incorporation of this knowledge in a text-to-speech system for Hindi. It includes a discussion on the properties of intonation patterns in Hindi such as declination/rising tendency, local fall-rise patterns in Hindi, resetting of  $F_0$  contour, significance of pause and on the effects of phonetic factors on intonation knowledge, and the issues in the incorporation of intonation knowledge in a text-to-speech system for Hindi.

**3.3a Properties of intonation knowledge in Hindi:** For the present analysis, we have used the reading style of speech. A corpus of 500 sentences was read out by two adult male native speakers of Hindi. Speech was digitized to 12 bits/sample at a sampling rate of 10 kHz. A 256-sample analysis frame with a shift of 64 samples was used for extracting pitch. The algorithms for pitch extraction are based on simplified inverse filter tracking (Markel 1972) and properties of group delay functions (Yegnanarayana et al 1991). In the following sections we discuss the properties of intonation patterns for continuous speech in Hindi in detail.

(i) *Declination/rising tendency* – Properties of  $F_0$  declination in Hindi can be summarized as follows: (1) Declination of  $F_0$  contour in Hindi is characterized by falls (valleys) and rises (peaks). (2) These falls and rises fluctuate between two abstract lines – a top line and a base line, drawn near or through all maxima and minima  $F_0$  values in a sentence, respectively. (3) The difference between valley and next peak (range of  $F_0$  contour) decreases with time. (4) In a neutral declarative sentence the maximum value of  $F_0$  will be located in the first *content word* (semantically meaningful word) itself. (5) In connected speech, the monosyllabic function words (words which have only grammatical value) conjoin with the preceding or following content words.

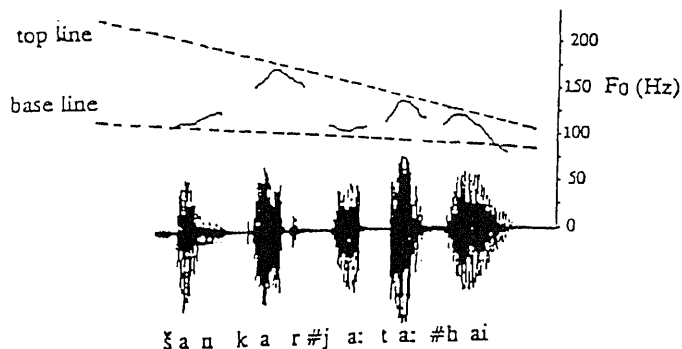


Figure 8. Speech waveform and  $F_0$  contour for simple declarative sentence /šankar ja:ta: hai/ 'Shankar goes'. # indicates word boundary.

Speech waveform and  $F_0$  contour for a natural utterance of a simple declarative sentence are shown in figure 8. The  $F_0$  contour starts at the initial syllable of the first word and rises towards the next target, that is, the final syllable of the first content word. The  $F_0$  rises and falls damp off at the end of the utterance. It is possible to draw a line connecting all the peaks (top line) and another connecting all the valleys (base line). Both lines decline monotonically and converge towards the end.

Interrogative sentences in Hindi can be broadly classified into two. They are (1) yes-no type questions and (2) question-word type questions. Yes-no type interrogative sentences in Hindi have the same grammatical structure as declarative sentences, except that optionally the question may include the question word /kya:/ usually at the beginning of the sentence. Question-word type interrogative sentences expect detailed answers and are marked in Hindi with any one of a set of interrogative words in Hindi. Intonation patterns for both types of interrogative sentences are different.

Questions expecting yes-no answers have a continuous rise in  $F_0$  contour and hence the top line and the base line rise towards the end. The intonation pattern for question-word type interrogative sentences exhibit a dual nature. The top line and the base line decline gradually up to the question word and then rise towards the end. The fall-rise pattern does not change with respect to the type of the sentence. Figure 9 shows the speech waveform and the  $F_0$  contour for a yes-no type interrogative

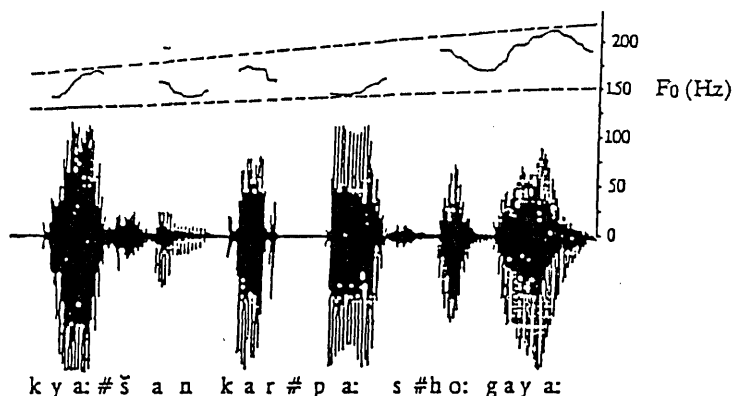
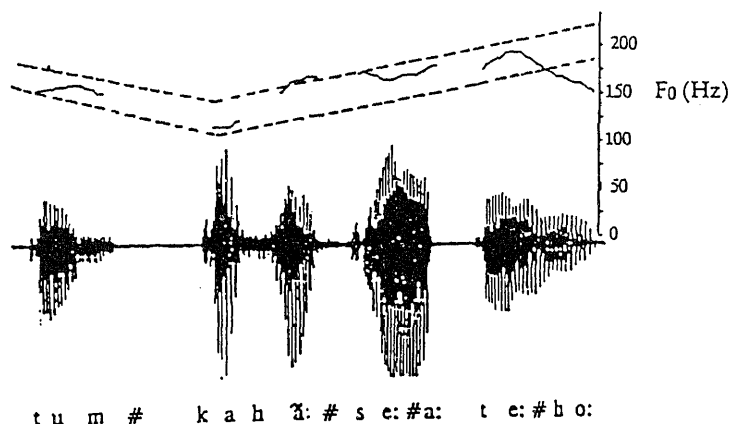


Figure 9. Speech waveform and  $F_0$  contour for a yes/no type interrogative sentence /kya: šankar pa:s ho:gaya:/ 'Has Shankar passed?'.



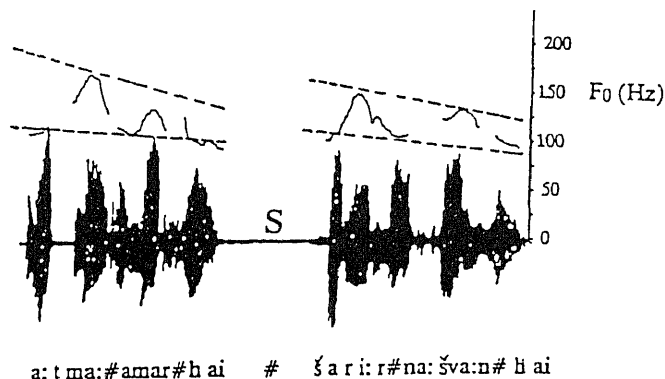
**Figure 10.** Speech waveform and  $F_0$  contour for a question-word type interrogative sentence /tum kəhā: se: a:te: ho:/ 'Where do you come from?'.

sentence. The  $F_0$  falls and rises repeatedly till the end of the utterance. Figure 10 shows the speech waveform  $F_0$  contour for a question word type interrogative sentence. Here the question word is /kahā:/ (second word in the sentence). The  $F_0$  contour declines up to the question word and then rises.

(ii) *Local fall-rise patterns in Hindi* –  $F_0$  contour of content words in Hindi exhibits a regular pattern of valleys and peaks, corresponding to the prominence of a particular syllable in a word or phrase. By analysing large amounts of data, we have observed some general features of valleys and peaks of  $F_0$  contour which are determined by the phonological pattern of the words. The following are some observations for Hindi sentences: (1) The valleys and peaks are mostly associated with the vowels which are the nuclei of the syllables. However, the exact target point (valley or peak) within the voiced region of the syllable is determined by several factors. For example, the peak of the nucleus get shifted to the coda (the consonant that follows the vowel nucleus of syllable) if the consonant is either nasal or lateral. (2) If the word is monosyllabic then the valley and the peak occur within the same syllable and hence  $F_0$  rises steadily. (3) In the case of disyllabic and trisyllabic words the peak occurs on the final syllable and the valley occurs on the initial syllable. (4) Tetrasyllabic words show two patterns: (a) a valley on the initial syllable and peak on the final syllable; (b) the valley and peak occur on alternate syllables and hence are characterized by two valleys and two peaks. The latter type is more likely in compound words. (5) The pattern for pentasyllabic words is similar to a combination of disyllabic and trisyllabic words.

(iii) *Resetting of  $F_0$  contour* –  $F_0$  resetting occurs across major syntactic boundaries and is accompanied by a pause. The part of utterance delimited by such a pause is called *intonational phrase*. The general properties of  $F_0$  resetting obtained from the analysis are summarized in the following sections.

The *initial peak  $F_0$*  ( $F_0$  value of the first peak of the first intonational phrase) is constant for a particular speaker. All other significant peaks and valleys in the subsequent clauses can be related to the initial peak  $F_0$ . Within a syntactic clause, the  $F_0$  contour is similar to the  $F_0$  contour of a simple sentence. That is, the declination



**Figure 11.** Speech waveform and resetting of  $F_0$  contour at clause boundary (S) in a complex declarative sentence /a:tma: amar hai śari:r na:śva:n hai/ 'Soul is immortal, body is mortal'.

of  $F_0$  contour is accompanied by local falls and rises. Figure 11 shows the effect of  $F_0$  resetting across syntactic boundaries. The sentence /a:tma: amar hai, śari:r na:śva:n hai/ (Soul is immortal, body is mortal) has two syntactic clauses and the  $F_0$  contour drifts down as a function of time till the occurrence of major syntactic break (at the end of /a:tma: amar hai/), which is also marked by a significant pause of duration of about 300 ms.

The major factors which can affect  $F_0$  resetting are physiological constraints, syntactic constraints and semantic constraints. Physiological constraints are the limitations imposed by the human speech production mechanism. Syntactic constraints include the changes in  $F_0$  resetting with respect to the changes in the type of the sentence. Semantic constraints are the semantic aspects which control the properties of  $F_0$  resetting.

(iv) *Significance of pause* – Pauses have been assigned to two main functions: (1) They separate large grammatical units, such as syntactic clauses. (2) They serve to clarify subgrouping of smaller units. Pauses can occur between words, intonational phrases and sentences. The characteristics of pauses for continuous speech in Hindi are summarized as below.

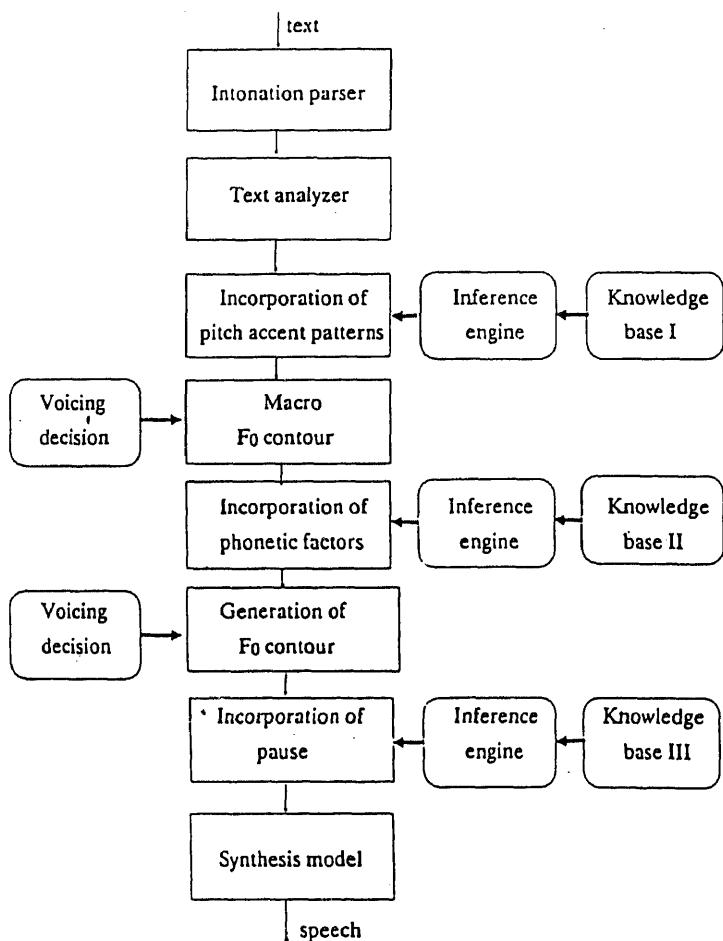
Speakers give different durations of pauses between words in continuous speech. The durations of pauses between words are controlled by several factors like lexical content of the words, position of the word in an intonational phrase and the phonetic factors of post-pause and pre-pause syllables. The amount of pause between words is much less than the amount of pause between intonational phrases. Between intonational phrases the amount of pause is determined by the type of constituent boundary. The amount of pause between sentences will be greater than pause between words and pause between intonational phrases.

(v) *Effect of segmental factors on  $F_0$  contour* – Acoustic phonetic behaviour of vowels and consonants of surrounding speech units alter the  $F_0$  values of vowels.  $F_0$  values of vowels were studied by embedding the test words in a carrier sentence. For this study we have selected several possible combinations of disyllabic words. The test words are mostly nonsense words where the vowel characteristics are studied for both

initial and final syllables separately. The results from this analysis are summarized below.

There is a correlation between height of the vowel and its inherent  $F_0$ . If other factors remain constant, high vowels (/i/ and /u/) exhibit higher  $F_0$  than low vowels (/a/). If the quantity of the vowel increases without changing any other factor, then inherent  $F_0$  also increases. For example, the long vowel /a:/ has higher  $F_0$  than the shorter counterpart /a/ at the same position. In all the cases, final vowels have greater  $F_0$  than initial vowels. Within each test word further regular variations in  $F_0$  were observed by changing the preceding or following syllables. However these changes are very small when compared with the changes due to other properties of  $F_0$  contour.

**3.3b Incorporation of intonation knowledge:** In a text-to-speech system, intonation refers to the periodicity of glottal pulse source in voiced speech segments. There are different stages required for the incorporation of intonation knowledge in a text-to-speech system. They are summarized as follows: (1) Input text has to be parsed to find out the type of sentence and the corresponding intonation behaviour. An intonation



**Figure 12.** Incorporation of intonation knowledge in a text-to-speech system for Hindi.

parser is used for the classification of sentences. The  $F_0$  contour changes with respect to the type of the sentence. (2) Text analysis has to be performed both at the word level and at the character level. Word-level analysis decides the importance of each word in the sentence. Character analyser determines the number of syllables in each word and classifies the syllables based on their acoustic phonetic behaviour. (3) Fall-rise patterns which include the decision of valleys and peaks have to be incorporated. After the incorporation of valleys and peaks a macro  $F_0$  contour is generated by joining successive valleys and peaks using a straight line based on the voiced/unvoiced classification of the corresponding basic units. (4) Segmental factors on the  $F_0$  contour have to be incorporated. The final  $F_0$  contour is generated at this stage using spline curves (Rogers & Adams 1989). (5) We have to incorporate the proper amounts of pauses between words, intonational phrases and sentences. During a pause, all source and system parameters are set to zero. (6) Intonation knowledge has to be represented using a suitable knowledge representation scheme in order to incorporate in a text-to-speech system. Our system is based on production system approach (Rich 1983). (7) Activation of intonation knowledge is achieved by means of a rule based inference engine with forward chain control strategy. Figure 12 places these issues in the overall scheme of our text-to-speech system.

The quality of synthetic speech obtained after the incorporation of intonation knowledge is tested for several sentences and compared with synthetic speech produced from the waveform concatenation model and the parameter concatenation model without intonation knowledge. The intelligibility and naturalness of synthetic speech increased significantly with the addition of intonation knowledge.

#### 4. Evaluation of the quality of synthetic speech

As the quality of the text-to-speech system improves, it is necessary to evaluate and compare the performance of each rule added to the system. Quality of synthetic speech is usually referred to the total auditory impression the listener experiences upon hearing the speech from the system. The listener's impression is influenced by the various constraints such as familiarity with the language, the inherent limitations of the human information system, the experience and training of the listener, the linguistic structure of the message set and the structure and the quality of the speech signal (Pisoni *et al* 1985; Childers & Ke Wu 1990).

There is no well-formed method for assessing the performance of synthetic speech quality. Assessment of the perceptual response by the human listener investigates the transmission of linguistic information from the speech signal and addresses specific questions such as how accurately synthetic characters and words are recognized, how well the meaning of synthetic utterance is understood and how easy it is to perceive and understand synthetic speech. Given below are discussions of some perceptual experiments done by us.

In order to test the improvement in quality, several sentences in Hindi were synthesized using (1) waveform concatenation model, (2) parameter concatenation model and (3) parameter concatenation model with the addition of different knowledge sources and demonstrated before several native and non-native speakers of Hindi. The results of these experiments are summarized in the following paragraphs.

From our experiments, we found that the speech obtained from parameter concatenation model (type 2) is better than the waveform concatenation model (type 1).

There are abrupt discontinuities in the synthetic speech obtained from type 1 at the boundaries of the basic units. This is removed by type 2 as concatenation is now done at the parameter level resulting in a reasonably smooth transition between boundaries. But there are various distortions in type 2 due to the lack of proper prosodic knowledge in synthetic speech.

The listeners were able to perceive the improvement in the quality of the synthetic speech in type 3, that is, in the parameter concatenation model after incorporating different knowledge sources. Depending on the outcome of the performance of the system, we can modify the system further, to attain the ultimate goal – to develop a text-to-speech system which is as good as natural.

## 5. Summary

In this paper, we have discussed the importance of knowledge sources in text-to-speech systems. Even though the emphasis was on Hindi, it can be extended to other Indian languages as well since the special features of Indian languages are considered in the design of the system. The main issues in the design and the development of such a system are the acquisition, representation and activation of knowledge at various levels, especially knowledge related to coarticulation, duration and intonation. All these knowledge sources are coded into a suitable form to incorporate into the system. The quality of speech from the text-to-speech system can be improved significantly with the addition of rules in the knowledge base. It is also essential to acquire knowledge related to other sources such as intensity variations and pauses between words and phrases.

## References

- Allen J 1985 A perspective of man-machine communication by speech. *Proc. IEEE* 73: 1541–1551
- Allen J, Hunnicutt M S, Klatt D H 1987 *From text-to-speech: the MITalk system* (Cambridge: University Press)
- Atal B S, Hanauer S L 1971 Speech analysis and synthesis by linear prediction of speech wave. *J. Acoust. Soc. Am.* 50: 637–655
- Childers D G, Ke Wu 1990 Quality of speech produced by analysis-synthesis. *Speech Commun.* 9: 97–117
- Fant G 1982 The voice source – acoustic modeling, Technical report, STL-QPSR 4/1982: 28–48
- Klatt D H 1976 Linguistic uses of segmental duration in English: acoustic and perceptual evidences. *J. Acoust. Soc. Am.* 60: 1208–1221
- Klatt D H 1980 Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* 67: 971–995
- Madhukumar A S, Rajendran S, Yegnanarayana B 1993 Intonation component of a text-to-speech system for Hindi. *Computer Speech and Language* 7: 283–301
- Makhoul J 1975 Linear prediction: a tutorial review. *Proc. IEEE* 63: 561–580
- Markel J D 1972 The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Acoust. Speech Signal Process.* 24: 399–418
- Ohman S E G 1966 Coarticulation in VCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.* 39: 151–168
- O'Shaughnessy D 1984 Design of a real-time French text-to-speech system. *Speech Commun.* 3: 233–243
- O'Shaughnessy D 1987 *Speech communication – Human and machine* (Reading, MA: Addison Wesley)

- Papamichalis P E 1987 *Practical approaches to speech coding* (Englewood Cliffs, NJ: Prentice Hall)
- Pisoni D B, Nusbaum H C, Green B G 1985 Perception of synthetic speech generated by rule. *Proc. IEEE* 73: 1665-1676
- Rajesh Kumar S R 1990 *Significance of durational knowledge in a text-to-speech system for Hindi*. M S Dissertation, Indian Institute of Technology, Madras
- Ramachandran V R, Yegnanarayana B 1992 Coarticulation rules for a text-to-speech system for Hindi. In *Proceedings of the Speech Technology Workshop*, Indian Institute of Technology, Madras, pp. 211-219
- Rich E 1983 *Artificial intelligence* (New York: McGraw Hill)
- Rogers D F, Adams J A 1989 *Mathematical elements for computer graphics* (New York: McGraw Hill)
- Yegnanarayana B, Murthy H A, Ramachandran V R 1991 Speech processing using modified group delay functions. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toronto, 2: 945-948
- Yegnanarayana B, Murthy H A, Sundar R, Alwar N, Ramachandran V R, Madhukumar A S, Rajendran S 1990 Development of a text-to-speech system for Indian languages. In *Frontiers of knowledge based computing systems* (eds) K M Rege, V P Bhatkar (Bombay: Narosa)





## Planning in bridge with thematic actions

DEEPAK KHEMANI

Department of Computer Science & Engineering, Indian Institute of Technology, Madras 600 036, India

E-mail: khemani@iitm.ernet.in

**Abstract.** The task of planning in a dynamic and an uncertain domain is considerably more challenging than in domains traditionally adopted by AI planning methods. Planning in real situations has to be a knowledge intensive process, particularly since it is not easy to predict all the effects of one's actions. Contract bridge offers a domain in which many of the issues involved in real world problems can be addressed without having to make simplifications in representation. Planning in the game of bridge takes us away from the traditional search-based methods (like the alpha-beta procedure), which are applicable in complete-information games like chess. In this paper we look at how knowledge can be structured to plan for declarer play in bridge. This involves deploying known move combinations, triggered by patterns which are abstracted out of the input, and then assembling the structures into a workable plan. The results demonstrate the viability of the proposed concepts.

**Keywords.** Knowledge-based planning; games; knowledge structures; uncertain environment.

### 1. Introduction

Traditionally, games invoke images of game-theoretic methods (Rapoport 1966). These methods essentially search in the game space to move towards the saddle point, which is the best one can achieve against a perfect opponent in a complete information situation. Our emphasis, on the other hand, is entirely on the use of knowledge for problem solving. Planning (Hendler 1990) has been one of the key areas of AI research. However, traditional AI planning methods make some very simplifying assumptions about the world (Marks & Hammond 1988). One such assumption is that the world is completely known and stable. The effect of any proposed actions can be predicted precisely. Thus one assumes that a totally correct plan can be generated, and can be executed perfectly. In practice, the world is neither completely known nor static. As a result planning is imperfect and plans may warrant changes *en route*.

Contract bridge provides an excellent domain for the exploration of knowledge-based methods in an uncertain environment (Khemani & Ramakrishna 1989). We

get an environment with all the advantages of working with games, viz. a discrete finite universe which does not require representational compromises, where symbolic interaction is complete, and where performance can be accurately measured. In addition, being a four-player incomplete-information game, bridge provides opportunities which two-person-zero-sum complete-information games like chess do not. These include the necessity to reason under uncertain conditions, the need for formal communication encoded in moves between players, and a gamut of problems to tackle (as opposed to the same starting board position in Chess and Go). Thus, looking for the right moves in bridge can be classified as a planning problem in a *hostile real world* (Noronha & Sarma 1991). At the same time bridge provides a large enough problem-solving environment which has held humankind spellbound for decades. A more detailed description of the problems in the game is given in appendix A.

We have observed that a search-based strategy is not suitable. Instead we adopt a knowledge-based approach in which the planner works with card combinations for which known techniques can be applied. Such knowledge is essentially heuristic in nature, aggregating the efforts of a deeper analysis involving various cases. Representational economy dictates that heuristic knowledge be expressed at a suitably abstract level. However, this means that the *use* of such knowledge by an autonomous system also requires the ability to move between the domain and the abstract levels. Consequently, the knowledge structures proposed in this paper are composed of three diverse generic components: recognition, analysis, and instantiation.

The granularity of the knowledge structures is such that they apply only to a combination of a few cards, rather than the play of the entire hand. That is, the heuristic knowledge, which is in the form of 'strong' structures, suggests partial plans. The play of most hands can be seen as a combination of such known plays. Selecting the different relevant structures, and combining their actions into a coherent plan is the task of a scheduler, which has a flavour of a 'weak' method.

Planning then, as described in the sections that follow, is accomplished as a two-stage process – the activation of various knowledge structures suggesting possible partial plans, followed by the assembling of a feasible plan using some of them.

## 2. Heuristic knowledge

Heuristic knowledge can be said to be knowledge that facilitates solving a class of problems. A body constituting such knowledge would in some sense "know" what to do in a given situation. By its nature, it is often domain-specific in nature, though one can think of abstract enough "rules" which would apply across many domains. Heuristic knowledge is essentially associative in nature, matching patterns at a suitably abstract level and suggesting a known solution method. For example, a person learning to brew tea may work with the following rule "when the water begins to boil, add the tea leaves and switch off the stove". To solve a problem, heuristic rules like these can be blindly employed without doing any deeper cause-effect analysis. This, however, would be only the first step towards expert behaviour. To be able to explain one's actions, and to be able to adapt to somewhat different situations, or even to break the rule, would require a deeper analysis. The next section illustrates this idea in more detail using the domain of bridge.

Assembling a heuristic knowledge base is, however, not a trivial task. We claim

that this is mainly because such knowledge cannot function in isolation\*. Instead, this knowledge applies to patterns in information which have been abstracted out of raw data, away from irrelevant detail. The process of abstraction may be methodologically quite different, but would have to work in tandem with the higher level reasoning, which we may call meta-semantic reasoning (Khemani 1988). In fact, the process of abstraction will be strongly influenced by the sort of features the meta-semantic heuristics need. On the other hand, solutions provided by the heuristic knowledge would again have to be effected at the domain level thus clearly defining their semantics. In this process the detail that was ignored during abstraction has to be again taken into account as actions have, in the end, to be visible at the domain level. Thus, one needs to incorporate diverse generic skills (Chandrasekaran 1986) in a system which is to function autonomously.

In addition, there is a need to structure the knowledge in a manner that relevant parts are invoked automatically in any situation. It has quite convincingly been shown in the context of understanding (Dyer 1986) that various types of knowledge structures are needed. These may vary from frames (Charniak 1978) and scripts (Schank & Abelson 1977) to handle stereotypical activity, to more complex structures (Lehnert 1987) to capture intentions, goals, plans and their myriad interrelations.

We represent planning knowledge for declarer play in bridge in the form of thematic acts (TA) (Khemani 1988). Each TA encapsulates a technique to tackle a recognizable situation, and the overall strategy is invoking one or more TA and weaving their suggested actions together into a single plan of action.

### 3. Knowledge for planning

Planning knowledge is required to charter a path through the combinatorially exploding number of possible sequences of low-level actions. Tactical knowledge to tackle each well-known situation can be represented at a higher level, somewhat like the macro operators introduced by Korf (1985). However, what is also needed is the ability to view the macro operators at a higher level, recognizing the situations in which they could be applicable and the changes that they may effect. We introduce a knowledge structure called thematic act which operates in an abstraction space. A TA has knowledge of patterns to look for in the raw data, knowledge of move combinations (a partial plan) which might be useful, and knowledge of the overall effect of the moves (which we call *meta-semantic* knowledge). The TA captures the theme behind all the similar move combinations which are required to achieve goals in similar situations. We explore the nature of such thematic knowledge in the context of contract bridge.

#### 3.1 Functional abstraction

In any interesting enough domain the number of different cases would be too many to put in memory individually. Instead the goal should be to try and extract general

---

\*Isolated heuristic systems, for example MYCIN (Shortliffe 1976), have been built of course but they are only prescriptive in nature. They capture the associations between syndrome (or the therapy), but the tasks of observing the symptoms and effecting are left to the human user.

principles which are applicable in many cases. Consider, for example, the concept of a 'finesse' in bridge. It can be illustrated by the following case, where all the cards are from the same suit, and "?" represents an unknown card<sup>+</sup>.

North: AQ		
West: ??	East: ??	(1)
South: 3 2		

South plays the 3. Let us say West plays the 8. If North plays the Q, then the play is said to be a finesse. If the K is in the West hand the North-South side makes an extra trick with the Q. If on the other hand the K is in the East hand, the finesse fails and East can win the trick.

The important feature in the above play is the combination of Ace & Queen which forms a 'tenace' over the West hand. Straightaway we can see that the South cards are insignificant, and could have been any of the small cards. Also we can see that the North hand could have had A K J, and one could have still finessed the Jack. What is more, if the A K Q and J had already been played, then a combination of 10 & 8 could have formed a tenace over East's 9. So *functionally* a tenace can be defined as follows

North: RC1 RC3		
South: x		(2)

which essentially says that if there is a rank-class-1 (RC1) card along with a rank-class-3 (RC3) card, and there is a small card (x) in the other hand, then one can *plan* to take a finesse against the opponent's RC2 card. The rank-class is different from the rank, in the sense that it clubs together cards of touching ranks held by a side. For example, if both the Ace and the King belong to the same side then both are of rank-class 1, whereas the rank of K is 2. In other words, since both the Ace and the King are held by the same side, the King could also be considered to be a winner.

Thus *the tenace is a functional abstraction* which characterizes the situation when a *finesse* can be played. Each card is viewed functionally by its role, which may vary in different situations, rather than by its identifier, which is constant. In addition one is looking for *combinations* which can be exploited. This is quite different from the accepted meaning of abstraction, for example in ABSTRIPS (Sacerdoti 1974), where the emphasis is only on ignoring unnecessary detail.

### 3.2 The level of knowledge

The knowledge captured in a TA is aimed at its use. It does not provide for a situation-action-effect analysis that could provide an explanation for *why* the recommended actions are useful. Consequently it cannot also cater to extra-thematic situations. Consider, for example, the following TA. The hands are

North: K 9 x		
South: A J xxx		(3)

We need four tricks from this combination, and can afford to lose one (but not two).

<sup>+</sup> Symbols pertaining to the game and used in the paper are defined in the appendix.

The standard play, which the TA recommends in this situation, is to play the A first, and then a small card towards the K-9, intending to play the 9. However, if West does not follow suit, win with the King, and play the 9 towards the Jack. It may be observed that this play is somewhat counter intuitive as it tends to disrupt the A-J tenace that exists.

The knowledge contained in the TA cannot provide explanations as to why this play is best. That requires a deeper analysis which deals with cases, and is along the following lines. There is a danger when one of the opponents holds Q-10-x-x. In such a situation the above play insures against losing two tricks. If West holds Q-10-x-x he is forced to play either the Q or the 10 lest North wins with the 9. And then the 9 and the Jack between them will restrict West to winning only one trick with the other card. If East holds the Q-10-x-x then West will show out on the second round. North wins, and now playing towards the Jack ensures that East can only win with the Queen. The intuitive play, on the other hand, with this card combination, is to first win the King, and then finesse the Jack. This play is best for winning five tricks, but when West has Q-10-x-x it loses two tricks. The recommended play for four tricks also guards against this eventuality. However, the ability to do the above analysis has no effect on the success of the TA in that situation.

There can be extra-thematic situations where the standard advice has to be ignored. One such situation arises when one is willing to alter one's risk-gain equation. Thus, it may happen, in the above example, that in a larger context (e.g. a tournament) one is willing to risk losing two tricks for the small gain of an extra trick (when only four are required). However, such reasoning is beyond the scope of this work.

#### 4. Planning with the TA

The goal in bridge is to arrive at a plan to make a certain number of tricks. The cards dealt define the starting position. The opponents' cards are not known. However some information may be available from the bidding, and more accrues as play proceeds. The next section looks at the role of this information in decision making. In this section we examine the role of TA in forming the plan.

The basic strategy used in the planning process is Means-Ends-Analysis (MEA) (Newell & Simon 1963). This applies easily to the bridge problem as the goal of making a certain number of tricks can be split into conjuncts of developing tricks in different suits. There are, however, overall constraints imposed due to sharing of common resources. For example, if we need to develop three tricks, and there exist two suits that can provide two each, it may still not be possible to develop both due to 'tempo' restrictions. So, the TA merely suggest partial plans to reduce a part or the whole of the difference, and it is left to an overall scheduler to assemble the best combination.

The TA provides the knowledge to tackle a given card combination in a given suit. It is not necessary, however, that a suit can be optimally played in isolation. For example, in the following card combination

North: A Q J  
South: 6 4 3

(4)

the recommended play is to finesse twice. This requires that the play be started twice from the South position. Each time the lead shifts to the North hand, since the trick

is won there. Thus, some other suit needs to provide an 'entry' into the South hand to finesse again. In this way the plays for different suits need to be interleaved and woven together. Notice also that the play for the above combination can be hierarchically composed using the finesse.

Each suit has some potential for developing a certain number of tricks with some probability. For example combination (3) can be played for 4 tricks with greater insurance than for 5 tricks, and it can be played for 3 tricks with assured results. However, there are different conditions required by each of the plays. The 3-trick play requires available *leeway* and *tempo* to be at least 2. These terms are defined below. The 4-trick play requires them to be at least 1, while the 5-trick play has no such constraints.

The selection of different TA combinations is dictated by some overall constraints. Constraints are of two types. One is the maximum number of tricks that the planner can afford to lose. For example, if you have contracted for 13 tricks, and need only 4 from the combination (3), you still cannot select the 4-trick play because it allows for losing one trick to the opponents. That is, the *leeway* required is one, but that available is zero. Therefore, even though the play suggests itself, the scheduler cannot select it.

A related constraint is that of *tempo*. This is determined by how close the opponents are to setting up tricks for themselves. It is measured in terms of the *number of times* the planner can lose tricks to the opponents before they are in a position to defeat the contract. It is the job of the scheduler to pick up a combination of thematic acts within the overall leeway and tempo constraints.

The *task* of planning is to combine thematic acts for different suit combinations to achieve the *overall* goal of tricks. In doing so one has to be careful that the different partial plans can be combined together to form a feasible line of action. For example, if combination (4) is replicated in all suits,

North: A Q J, A Q J, A Q J, A Q J 10  
South: 6 4 3, 6 4 3, 6 4 3, 6 4 3 2 (5)

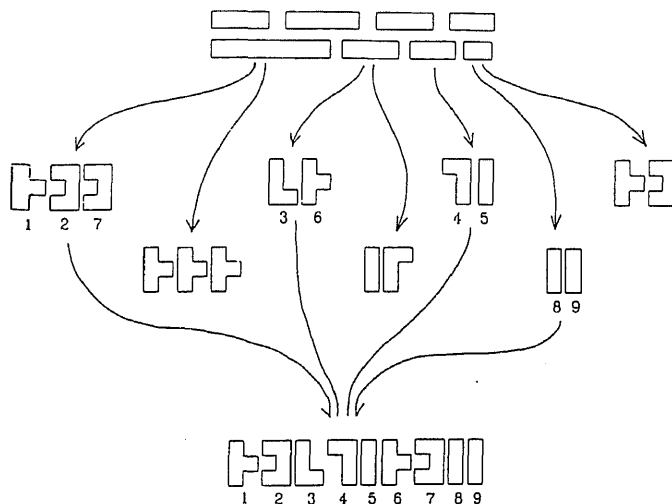
then not more than one finesse could be possible, though the TA would create partial plans for 9 of them! However, if some of the suits were interchanged,

North: A Q J, 6 4 3, A Q J, 6 4 3 2  
South: 6 4 3, A Q J, 6 4 3, A Q J 10 (6)

then all 9 finesses would be effected, possibly winning all 13 tricks.

Hence by doing a meta-semantic analysis of the plays offered by different suits a suitable plan can be arrived at. The thematic acts themselves are tuned to the best suit play, i.e., one that has highest probability of success. The task now is to combine some of them into forming a plan. In doing so one may discover that a TA cannot be directly applied, and one may have to look for an alternate play which can fit into the plan. The plan generated will have a certain probability of success that is determined by the probability of the assumptions made *en route* being true. By and large the best plan selects itself, since the acts take care to suggest the best individual play. But if the best individual thematic acts cannot be combined to form a plan one may have to look at various options. This is discussed in § 5. Also, sometimes the mounting evidence may require meta-thematic analysis for selecting a plan.

It can be seen that there are two distinct aspects to the task of planning. The first,



### ANNOTATION

The larger problem contains a set of smaller problems for which (partial) solutions are known.

A thematic planner looks at subproblems and the thematic acts suggest partial plans. Numbers below show their position in the final plan.

A general purpose scheduler looks at various thematic actions and assembles together a coherent plan of actions which mesh together.

Figure 1. Two-stage planning - A schematic flowchart.

which involves applying known solutions to recognizable parts of the problem, is more knowledge intensive and has a 'strong' flavour. The TA provide this capacity, and each partial plan is retrieved complete with its own ordering (Berlin 1985). On the other hand, the scheduler which assembles plans is essentially the embodiment of a weak method. Figure 1 schematically illustrates the formation of partial plans and the assembling of a total plan with some of them. Such an architecture is imperative if planning is to be done in any complex domain. An autonomous agent in such a domain will have to cater to many goals which crop up as it functions. Planning, therefore, can no longer be viewed as problem-solving for a well defined goal, but is an ongoing process. A basic requirement then, in such a domain, is the ability to incorporate partial solutions in one's activity. This will also enable opportunistic planning (Hayes-Roth & Hayes-Roth 1979) specifically while executing plans (Birnbaum 1985).

## 5. Choosing between plans

A given plan succeeds when the implied assumptions it makes are satisfied. One of the strategies in planning is to make as few assumptions as possible. The assumptions made select a subset  $S1$  of the all possible situations  $U$  when the given plan will succeed (figure 2).

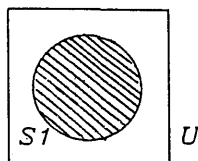


Figure 2. The subset  $S1$  of the situation when the plan works in the universe  $U$ .

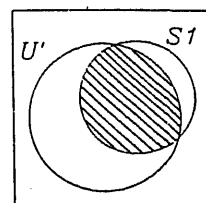


Figure 3. The subset  $S1' = S1 \cap U'$  in the constrained universe  $U'$ .



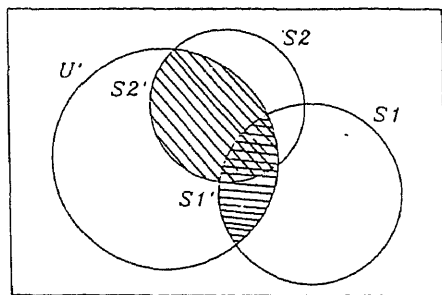


Figure 4. Sets  $S1'$  and  $S2'$  for plans  $P1$  and  $P2$  in the constrained universe  $U'$ .

The aim in planning with thematic acts is to maximize the size of  $S1$ . However, as play proceeds, one may get more information about the actual situation. This contracts the universe  $U$  to a set  $U'$ . The subset  $S1' = S1 \cap U'$ , now is the set of situations when our plan succeeds (figure 3).

Between two competing plans  $P1$  and  $P2$ , the sets  $S1'$  and  $S2'$  now define the probability of success. The shaded area in figure 4 shows the differing cases for which  $P1$  and  $P2$  succeed, given  $U'$ . It can clearly be seen that the set  $U'$  can play a decisive role in selecting the better plan. For example, without any knowledge about world  $U$ , plan  $P1$  is a better plan if  $S1$  is larger than  $S2$ . However knowing that the universe is constrained to  $U'$  plan  $P2$  turns out to be better if  $S2'$  is larger than  $S1'$ .

The overall strategy in planning to play a bridge hand is two-fold. One, to try and synthesize plans which make as few assumptions as possible. This is equivalent to saying that the plan should work in as many situations as possible. The second aspect of planning is to try and glean more information about the world. The effort here is to try and reduce the number of situations that have to be catered for. One can see that such a strategy is also likely to prove useful in any incomplete information domain.

## 6. Implementation

We have implemented a program for declarer play in bridge. The input to the program is the declarer's hand (South), the dummy (North), and the opening lead made by West. Currently the program plays only no trump (NT) contracts. The overall task is broken up into the following modules.

The program begins by abstracting the relevant patterns out of the cards. It counts its tricks, and determines the goal (i.e. the number of tricks that have to be generated), and the constraints. For example, when planning for a 7NT contract there is no leeway, but a 3NT contract allows four tricks for the opponents and the planner often has some tempo. The program then looks for any card combinations that it 'knows' to tackle effectively. This information is provided by the corresponding TA and the pattern also serves to index the TA. The planner constructs partial plans using the procedural knowledge contained in the TA. The current implementation generates one partial plan for each suit and passes it on to the scheduler. If this cannot fit into the overall plan, then the program backtracks and tries another. The ordering of the actions suggested by the TA is done in a heuristic manner by the scheduler. Thus the right play in the case referred to in Berlin (1985) is found directly because of the heuristics employed, rather than by a deeper analysis of interaction of subgoals. The

scheduler orders the actions to link up the different thematic acts. For example, combination (4) requires that the suit be played from the South hand, and it is the job of the scheduler to incorporate the play when the opportunity presents itself. During play, if any assumptions are violated then control is passed back to the planner. This also happens when an unexpected trick comes up. A more detailed description of the program is given in Ramachandran & Khemani (1992).

## 6.1 Results

The trace left by the program as it plans and plays a few hands is included in appendix B. The trace constitutes of canned messages printed by various sections when they execute. The input to the program are the North and South hands. For our convenience the East and West hands have also been included in the printout.

In hand 1, needing to make all 13 tricks the program spots the tenaces in Hearts. It observes that it can cater to the worst break by finessing 3 times, and forms a plan accordingly. In hand 2 the program relies on the spade break for tricks, and noting the shortage of entries, plans to 'duck' one round of the suit. Hand 3 is a sort of combination of 1 and 2. The feature of using a delayed finesse to solve the communication problem has been encoded as a separate TA. Hands 4 and 5 are the same for North and South (thus leading to the same plan), but different for East and West. In both cases the assumed spade break does not materialize, and the program looks at alternative sources. In hand 4 it avoids the spade finesse because it knows that it is bound to fail, while in 5 it chooses it knowing that it will succeed.

## 7. Conclusions

In this paper we have looked at a two-stage mechanism for planning, using contract bridge as the domain. In the first stage partial plans are suggested by a theme-based planner. The planner employs knowledge structures called thematic acts which have the flavour of strong AI. Each TA encapsulates the knowledge required for solving a problem in a familiar situation in a well-defined way. To retrieve and use this knowledge in applicable situations, the theme-based planner requires three kinds of processing. First, in functional abstraction the patterns which serve as indices to the TA are extracted. Then, meta-semantic reasoning computes the cause-effect relations of the TA in the situation. Finally, procedural knowledge orchestrates the actual domain level primitive actions. The partial plans thus are combinations of primitive actions, not necessarily complete in themselves, that can be used to solve a part of the problem.

In the second stage, a scheduler combines the actions suggested by the TA into a coherent plan. This may often require that actions of different TA be interleaved. The scheduler is basically an embodiment of a weak method, but one which operates under the constraints imposed by the first stage of thematic planning. In our implementation the strategy of means-ends-analysis has been adopted, as it is well-suited to the domain of planning in bridge.

The results of our implementation have been very encouraging, as can be witnessed in the examples shown. We feel that such a scheme of planning is particularly attractive when planning in a complex domain where (partial) plans for newer goals have to be incorporated into the larger plans of an agent who may be continuously interacting with the world.

## Appendix A. Contract bridge: The problems

In this appendix we look at the game of contract bridge, highlighting some of the problems that are of interest. We observe that unlike chess, which is a two-person zero-sum complete-information game, bridge cannot be tackled by elegant mechanisms like the minimax search method. This is because bridge is *not* a complete information game. Then, since one does not know the cards held by the opponents, one cannot project the play into the future to try and discover which strategy is most profitable. Instead one has to rely on some knowledge-intensive method. Bridge can be classified as a two-side incomplete-information game. Further complication is introduced by the fact that each side constitutes two persons. Therefore communication is vital. Not only does one have to convey information, within the rules of the game of course, to the partner but one also needs to intercept opponents messages to learn their intentions. Almost as a corollary, at a more sophisticated level one may even want to send out misleading signals to lead opponents astray.

Contract bridge is played with a regular pack of 52 cards dealt randomly and equally among 4 players. Let us call them North, South, East and West, according to their position on the table. North and South are partners, as are East and West. The cards are ranked in the order Ace (A), King (K), Queen (Q), Jack (J), 10 (T), 9, ... 2 in each suit. Each player plays a card, in clockwise order, and the highest ranking card wins the trick. Thirteen such tricks are played, each time the winner of the precious trick starting play. This constitutes *one deal or one hand*.

There are two stages of play in each deal, viz. bidding, followed by the play of cards. The goal in a deal is to maximize *points*. The points essentially depend upon bidding. Bids are made for the number of tricks the side promises to make, given the stated 'trump' suit. Eventually the highest bid is accepted in each deal. This is known as the *contract*. Generally, the higher a side bids the more points it is likely to win, *provided* it can fulfill the contract. That is, if the side can make the number of tricks it has bid for. If it succeeds, it wins some points. Let us call them success-points. If it loses, then the opponents get some points instead, which we can call penalty-points.

### A1. Bidding

The straightforward goal in bidding is to bid the *highest* number of tricks one thinks one *can* make. That is, to maximize success-points won. The means used in this process are the following,

- (1) Evaluation of own hand.
- (2) Communication with partner.
- (3) Projection of play.

Of these, the first two are simpler and can possibly be handled by heuristic methods. The third is more difficult, as it would involve constructing plausible distributions (based on the bids heard, and on probability) and then projecting the play.

A more complex goal is to make a *sacrifice bid*. This essentially means overbidding, over an opponent bid, with the hope that the penalty-points loss will be lesser than the opponents' expected success-points gain, thus being an overall gain.

Even more complex goals are to sabotage the opponents communication. This may mean consuming the bidding space (jamming the communications channel), or even making 'false' bids to confuse opponents. In the process, an enterprising planner

may make an 'advance sacrifice' to 'push' the opponents higher than they can manage, or to escape with a lighter penalty.

Considering that all these processes happen when the planner can see only one hand, one observes that bidding is probably the more difficult part of the game.

## A2. *Play of the hand*

Once bidding is over, the goal for the play stage has been defined. One side has the contract, and is required to make the bid number of tricks. At this stage one player of the contracting side (called the dummy) exposes the cards to everybody, while the other (called the declarer) plans and executes the play. The opposing side (called defenders) are said to *defend* the contract, trying in fact to defeat the accomplishment of the contract by the declarer.

One can see that the situation at this stage is non-symmetric. The declarer knows the entire strength of his side, and is in total control of the play of the cards. He is also aware of the entire assets of the defence, in terms of material strength, since they have the remaining 26 cards. Each defender knows only his own hand, and cannot see his partner's hand. Therefore the two defenders have to combine their efforts to try and achieve the goal. This necessarily involves (formal) communication between the two. Both can see the dummy also.

Since the cards of all the players cannot be seen, one cannot project moves into the future. Methods like minimax search are therefore ruled out immediately. Instead, the success of a strategy can only be estimated based on the probabilistic distribution of cards, and any information gleaned from the communication taking place. The strategies themselves are derived from knowledge about the various known methods of tackling various card combinations. A constant endeavour is to cater to as many situations as possible, to maximize the chances of success.

The straightforward goal in the play of the hand is to make the number of tricks as stated in the contract. The emphasis is on maximizing the probability of success. If success is assured, then the goal can be revised to increasing the number of tricks won, as some more points can then be gained. If success seems unlikely, then a planner may even choose to minimize losses, i.e. the penalty-points won by the opponents. Like in bidding, the planner may attempt to do better than par, by exploiting the incomplete information that the opponents have. This may introduce complex 'meta-level' goals of protecting information, or sending out misleading signals.

Thus, we see that unlike in games like chess, where a clear-cut strategy of aiming for the minimax value (saddle) points is meaningful, in bridge one has to largely grapple with incomplete information. In the face of such uncertainty, planning in the game of bridge can only be a complex knowledge-intensive activity.

## Appendix B. *Traces of the program*

This appendix contains the traces left by the program as it plans and plays some hands. The trace contains the complete hand, followed by the trace, followed by the record of play. Even in cases where the system replans after some play, the organization of text remains the same. Text in square brackets is comments by the author.

## Example B1

[Hand 1]<sup>†</sup>

Contract = 7 Notrump

North – Dummy

♠ 752  
♥ AQJT5  
♦ A43  
♣ K3

West

♠  
♥ K976  
♦ T982  
♣ J964

Opening lead – J ♦

East

♠ JT98643  
♥  
♦ 765  
♣ T87

South

♠ AKQ  
♥ 8432  
♦ KQ  
♣ AQ52

*The planning phase...*

Have 10 top tricks... Need 3 tricks more.... Tempo available 10. If Hearts break 4–0 we get 3 extra tricks losing 1... can't afford that... first glance over.... Tenace over West found in Hearts. Have planned for 3 tricks in Spades, 5 tricks in Hearts, 3 tricks in Diamonds and 3 tricks in Clubs. [Note that it was planned for  $5 + 3 + 3 + 3 = 14$  tricks. It does not need to decide which one of the winners it will have to discard. The discarding problem has been passed over to the scheduler which will finally throw away a winner.]

*The play..**Examining lead... Looks like top of a sequence.*

(A \* indicates that the corresponding player had the lead.)

No.	West	North	East	South	Won by
1	*J ♦	3 ♦	5 ♦	K ♦	South
2	2 ♦	2 ♠	3 ♠	*A ♠	South
3	6 ♥	T ♥	4 ♠	*2 ♥	North
4	4 ♣	*5 ♠	6 ♠	K ♠	South
5	7 ♥	J ♥	6 ♦	*3 ♥	North
6	8 ♦	*7 ♠	8 ♠	Q ♠	South
7	9 ♥	Q ♥	9 ♠	*4 ♥	North

<sup>†</sup>The symbols ♠, ♥, ♦ and ♣ are used for the suits spades, hearts, diamonds & clubs, respectively.

8	K♥	*A♥	7♦	8♥	North
9	9♦	*5♥	7♣	5♣	North
10	T♦	*4♦	T♣	Q♦	South
11	6♣	K♣	8♣	*2♣	North
12	9♣	*A♦	T♣	Q♣	North
13	J♣	*3♣	J♣	A♣	South

Tricks made = 13

Tricks contracted = 13

Good one! Made the contract!!!

### Example B2

[Hand 2]

Contract = 3 Notrump

North – Dummy

♠ A 7 6 4 3

♥ 7 6 5

♦ 5

♣ 8 7 5 4

West

♠ T 8

♥ Q J T 2

♦ J T 9

♣ T 8

Opening lead – Q♦

East

♠ Q J 2

♥ 8 4 3

♦ 8 7 6 3

♣ K J 6

South

♠ K 9 5

♥ A K 9

♦ A K 4 2

♣ A 3 2

### The planning phase...

Have 7 top tricks. ... Need 2 tricks more. ... Tempo available 4. If Spades break 3–2 we get 2 extra tricks losing 1. If Clubs break 3–3 we get 1 extra trick losing 2. ... first glance over... playing low to low due to lack of entries. Plan workable.... Have planned for 4 tricks in Spades, 2 tricks in Hearts, 2 tricks in Diamonds and 1 trick in Clubs.

The play..

(A \* indicates that the corresponding player had the lead.)

No.	West	North	East	South	Won by
1	*Q ♦	5 ♦	3 ♦	A ♦	South
2	8 ♠	3 ♠	2 ♠	*5 ♠	West
3	*J ♦	7 ♥	6 ♦	K ♦	South
4	T ♠	4 ♠	J ♠	*K ♠	South
5	2 ♥	A ♠	Q ♠	*9 ♠	North
6	9 ♦	*7 ♠	3 ♥	9 ♥	North
7	T ♦	*6 ♠	4 ♥	2 ♦	North
8	T ♥	*5 ♥	8 ♥	A ♥	South
9	Q ♥	6 ♥	6 ♠	*K ♥	South
10	9 ♠	4 ♠	J ♠	*A ♠	South
11	T ♠	5 ♠	8 ♦	*4 ♦	East
12	J ♥	7 ♠	*7 ♦	2 ♠	East
13	Q ♠	8 ♠	*K ♠	3 ♠	East

Tricks made = 9

Tricks contracted = 9

Good one! Made the contract!!!

Example B3

[Hand 3]

Contract = 7 Notrump

North - Dummy

♦ A Q 6 5 4 3

♥ J 7

♦ 8 4

♣ T 6 5

West

♦ K J 9

♥ Q 9 8

♦ Q 9 6

♣ Q 8 3

Opening lead - K ♦

East

♦ T 7

♥ T 6 3 2

♦ T 5 2

♣ J 9 7 2

South

♦ 8 2

♥ A K 5 4

♦ A J 7 3

♣ A K 4

The planning phase...

Have 6 top tricks... Need 3 tricks more... Tempo available 4. If Spades break 3-2 we get 3 extra tricks losing 2... first glance over... playing low to low due to lack of

entries... plan workable.... Have planned for 5 tricks in Spades, 2 tricks in Hearts, 1 trick in Diamonds and 2 tricks in Clubs.

*Examining lead*... Looks like top of nothing lead.

*The play...*

(A \* indicates that the corresponding player had the lead.)

No.	West	North	East	South	Won by
1	*K ♦	4 ♦	2 ♦	3 ♦	West
2	*3 ♣	5 ♣	J ♣	A ♣	South
3	9 ♣	3 ♣	T ♣	*8 ♣	East
4	8 ♣	6 ♣	*2 ♣	K ♣	South
5	J ♣	Q ♣	7 ♣	*2 ♣	North
6	K ♣	*A ♣	2 ♥	4 ♥	North
7	8 ♥	*6 ♣	3 ♥	5 ♥	North
8	6 ♦	*5 ♣	6 ♥	7 ♦	North
9	9 ♥	*4 ♣	T ♥	J ♦	North
10	Q ♥	*7 ♥	7 ♣	A ♥	South
11	9 ♦	J ♥	5 ♦	*K ♥	South
12	Q ♦	8 ♦	T ♦	*A ♦	South
13	Q ♣	T ♣	9 ♣	*4 ♣	West

Tricks made = 10

Tricks contracted = 9

Good one! Made the contract!!!

*Example B4*

[Hand 4]

Contract = 7 Notrump

*North – Dummy*

♦ AK32  
♥ AQT98  
♦ J32  
♣ Q

*West*

♦ J975  
♥ K3  
♦ 7  
♣ T7652

Opening lead – J ♣

*East*

♦ 4  
♥ J6542  
♦ T965  
♣ 943

*South*

♦ QT86  
♥ 7  
♦ AKQ84  
♣ AK8



*The planning phase...*

Have 11 top tricks.... Need 2 tricks more.... Tempo available 0. If Spades break 3-2 we get 1 extra trick losing 0. If Hearts break 6-1 we get 2 extra tricks losing 2... can't afford that. If Diamonds break 4-1 we get 1 extra trick losing 0... first glance over....

[Note that it does not look for spades as it knows that the break is not favourable.]

Have planned for 4 tricks in Spades, 1 trick in Hearts, 5 tricks in Diamonds and 3 tricks in Clubs.

*Examining lead...* Looks like top of a sequence.

*After trick 3*

Plan aborted because of Expectation Failure.

*New plan being generated...*

Have 8 top tricks.... Need 2 tricks more.... Tempo available 0. If Hearts break 6-1 we get 2 extra tricks losing 2... can't afford that. If Diamonds break 4-1 we get 1 extra trick losing 0... first glance over.... Tenace over East found in Spades. Finesse not possible as opponent has shown out. Tenace over West found in Hearts. Have planned for 1 trick in Spades, 2 tricks in Hearts, 5 tricks in Diamonds and 2 tricks in Clubs.

*The play...*

(A \* indicates that the corresponding player had the lead.)

---

No.	West	North	East	South	Won by
1	*J ♣	Q ♣	3 ♣	8 ♣	North
2	5 ♣	*A ♣	4 ♣	6 ♣	North
3	7 ♣	*K ♣	4 ♣	8 ♣	North
4	9 ♣	*2 ♣	9 ♣	Q ♣	South
5	2 ♣	9 ♥	2 ♥	*A ♣	South
6	5 ♣	8 ♥	5 ♦	*K ♣	South
7	3 ♥	Q ♥	4 ♥	*7 ♥	North
8	K ♥	*A ♥	5 ♥	T ♣	North
9	7 ♦	*J ♦	6 ♦	4 ♦	North
10	6 ♣	*2 ♦	9 ♦	A ♦	South
11	7 ♣	3 ♦	T ♦	*K ♦	South
12	T ♣	3 ♣	6 ♥	*Q ♦	South
13	J ♣	T ♥	J ♥	*8 ♦	South

---

Tricks made = 13

Tricks contracted = 13

Good one! Made the contract!!!

## Example B5

[Hand 5]

Contract = 7 Notrump

North – Dummy

♠ A K 3 2  
 ♥ A Q T 9 8  
 ♦ J 3 2  
 ♣ Q

West

♠ 4  
 ♥ J 6 5 4 2  
 ♦ 7  
 ♣ T 7 6 5 2

Opening lead – J ♣

East

♠ J 9 5 3  
 ♥ K 3  
 ♦ T 9 6 5  
 ♣ 9 4 3

South

♠ Q T 8 6  
 ♥ 7  
 ♦ A K Q 8 4  
 ♣ A K 8

*The planning phase...*

Have 11 top tricks... Need 2 tricks more... Tempo available 0. If Spades break 3–2 we get 1 extra trick losing 0. If Hearts break 6–1 we get 2 extra tricks losing 2... can't afford that. If Diamonds break 4–1 we get 1 extra trick losing 0... first glance over... Have planned for 11 tricks. Have planned for 4 tricks in Spades, 1 trick in Hearts, 5 tricks in Diamonds and 3 tricks in Clubs.

*The play...*

*Examining lead*... Looks like top of a sequence.

*After trick 3*

Plan aborted because of Expectation Failure

*New plan being generated...*

Have 8 top tricks... Need 2 tricks more... Tempo available 0. If Hearts break 6–1 we get 2 extra tricks losing 2... can't afford that. If Diamonds break 4–1 we get 1 extra trick losing 0... first glance over... Have planned for 8 tricks. Tenace over East found in Spades, Have placed the necessary card with opponent. Sure finesse!! Tenace over West found in Hearts, Have planned for 2 tricks in Spades, 1 trick in Hearts, 5 tricks in Diamonds and 2 tricks in Clubs.

[Note that the finesse in hearts is abandoned in preference to the finesse in spades as 10 tricks are available without it.]

The play...

(A \* indicates that the corresponding player had the lead.)

No.	West	North	East	South	Won by
1	*J ♣	Q ♣	3 ♣	8 ♣	North
2	4 ♣	*A ♣	5 ♣	6 ♣	North
3	2 ♣	*K ♣	7 ♣	8 ♣	North
4	2 ♥	*A ♥	3 ♥	7 ♥	North
5	5 ♣	*2 ♣	9 ♣	T ♣	South
6	6 ♣	8 ♥	4 ♣	*A ♣	South
7	7 ♣	Q ♥	9 ♣	*K ♣	South
8	7 ♦	3 ♣	J ♣	*Q ♣	South
9	T ♣	J ♦	5 ♦	*4 ♦	North
10	4 ♥	*2 ♦	6 ♦	A ♦	South
11	5 ♥	3 ♦	9 ♦	*K ♦	South
12	6 ♥	T ♥	T ♦	*Q ♦	South
13	J ♥	9 ♥	K ♥	*8 ♦	South

Tricks made = 13

Tricks contracted = 13. Good one! Made the contract!!!

## References

- Berlin D L S 1985 SPAN: Integrated problem solving tactics. *Proc. IJCAI-85* p 1047
- Birnbaum L 1985 A short note on opportunistic planning and memory in arguments. *Proc. IJCAI-85* p. 281
- Chandrasekaran B 1986 Generic tasks in knowledge-based reasoning: High-level building blocks for expert system design. *IEEE Expert* 1: 23-30
- Charniak E 1978 On the use of framed knowledge in language comprehension. *Artif. Intell.* 11: 225-265
- Dyer M G 1986 *In-depth understanding* (Cambridge: MIT Press)
- Hayes-Roth B, Hayes-Roth F A 1979 A cognitive model of planning. *Cognitive Sci.* 3: 275-310
- Hendler J, Tate A, Drummond M 1990 AI planning: Systems and techniques. *AI Mag.* Summer: 61-77
- Korf R E 1985 *Learning to solve problems by searching for macro-operators* (Boston: Pitman)
- Khemani D 1988 *Theme based planning in an uncertain environment*. Ph D thesis, Department of Computer Science & Engineering, Indian Institute of Technology
- Khemani D, Ramakrishna R S 1989 Bridge: A benchmark for knowledge based planning. *Commun. Cogn. Artif. Intell. (CC-AI)* 6: 137-151
- Lehnert W G 1987 Knowledge-based natural language understanding. In *Exploring artificial intelligence - Survey talks from the National Conference on Artificial Intelligence - 1986 & 1987* (ed.) H E Shrobe (San Mateo: Morgan Kaufmann)
- Marks M, Hammond K J, Converse T 1988 Planning in an open world: A pluralistic approach. *Proceedings of Workshop on Case-Based Reasoning*
- Newell A, Simon H A 1963 GPS: A program that simulates human thought. In *Computers and thought*. (eds) E A Feigenbaum, J Feldman, (New York: McGraw-Hill)
- Noronha S J, Sarma V V S 1991 Knowledge-based approaches for scheduling problems: A survey. *IEEE Trans. Knowledge Data Eng.* 3: 160-171
- Ramachandran R, Khemani D 1992 Planning declarer play in bridge, Tech. Rep. IITM-CSE-92-02
- Rapoport A 1966 *Two-person game theory* (Ann Arbor: Univ. of Michigan Press)
- Sacerdoti E 1974 Planning in a hierarchy of abstraction spaces. *Artif. Intell.* 5: 115-135
- Schank R C, Abelson R 1977 *Scripts, plans, goals, and understanding* (Hillsdale, NJ: Lawrence Erlbaum)
- Shortliffe E H 1976 *Computer-based medical consultations: MYCIN* (New York: Elsevier)

## Artificial neural networks for pattern recognition

B YEGNANARAYANA

Department of Computer Science and Engineering, Indian Institute of Technology, Madras 600 036, India

E-mail: yegna@iitm.ernet.in

MS received 12 April 1993; revised 8 September 1993

**Abstract.** This tutorial article deals with the basics of artificial neural networks (ANN) and their applications in pattern recognition. ANN can be viewed as computing models inspired by the structure and function of the biological neural network. These models are expected to deal with problem solving in a manner different from conventional computing. A distinction is made between pattern and data to emphasize the need for developing pattern processing systems to address pattern recognition tasks. After introducing the basic principles of ANN, some fundamental networks are examined in detail for their ability to solve simple pattern recognition tasks. These fundamental networks together with the principles of ANN will lead to the development of architectures for complex pattern recognition tasks. A few popular architectures are described to illustrate the need to develop an architecture specific to a given pattern recognition problem. Finally several issues that still need to be addressed to solve practical problems using ANN approach are discussed.

**Keywords.** Artificial neural network; pattern recognition; biological neural network.

### 1. Introduction

Human problem solving is basically a pattern processing problem and not a data processing problem. In any pattern recognition task humans perceive patterns in the input data and manipulate the pattern directly. In this paper we discuss attempts at developing computing models based on artificial neural networks (ANN) to deal with various pattern recognition situations in real life.

Search for new models of computing is motivated by our quest to solve natural (intelligent) tasks by exploiting the developments in computer technology (Marcus & van Dam 1991). The developments in artificial intelligence (AI) appeared promising till a few years ago. But when the AI methods were applied to natural tasks such as in speech, vision and natural language processing, the inadequacies of the methods

showed up. Like conventional algorithms, AI methods also need a clear specification of the problem, and mapping of the problem into a form suitable for the methods to be applicable. For example, in order to apply heuristic search methods, one needs to map the problem as a search problem. Likewise, to solve a problem using a rule-based approach, it is necessary to explicitly state the rules governing it. Scientists are hoping that computing models inspired by biological neural networks may provide new directions to solving problems arising in natural tasks. In particular, it is hoped that neural networks would extract the relevant features from input data and perform the pattern recognition task by learning from examples, without explicitly stating the rules for performing the task.

The objective of this tutorial paper is to present an overview of the current approaches based on artificial neural networks for solving various pattern recognition tasks. From the overview it will be evident that the current approaches still fall far short of our expectations, and there is scope for evolving better models inspired by the principles of operation of our biological neural network. This paper is organized as follows: In § 2 we discuss the nature of patterns and pattern recognition tasks that we encounter in our daily life. We make a distinction between pattern and data, and also between understanding and recognition. In this section we also briefly discuss methods available for dealing with pattern recognition tasks, and make a case for new models of computing based on artificial neural networks. The basics of artificial neural networks are presented in § 3, including a brief discussion on the operation of a biological neural network, models of neuron and the neuronal activation and synaptic dynamics. Section 4 deals with the subject matter of this paper, namely, the use of principles of artificial neural networks to solve simple pattern recognition tasks. This section introduces the fundamental neural networks that laid the foundation for developing new architectures. In § 5 we discuss a few architectures for complex pattern recognition tasks. In the final section we discuss several issues that need to be addressed to develop artificial neural network models for solving practical problems.

## **2. Patterns and pattern recognition tasks**

### *2.1 Notion of intelligence*

The current usage of the terms like AI systems, intelligent systems, knowledge-based systems, expert systems etc., are intended to show the urge to build machines that can demonstrate intelligence similar to human beings in performing some simple tasks. In these tasks we look at the performance of a machine and compare it with the performance of a person. We attribute intelligence to the machine if the performances match. But the way the tasks are performed by a machine and by a human being are basically different; the machine performing the task in a step-by-step sequential manner dictated by an algorithm, modified by some known heuristics.

The algorithm and the heuristics have to be derived for a given task. Once derived, they generally remain fixed. Typically, implementation of these tasks requires large number of operations (arithmetic and logical) and also a large amount of memory. The trends in computing clearly demonstrate the machine's ability to handle a large number of operations (Marcus & van Dam 1991).

## 2.2 Patterns and data

However, the mere ability of a machine to perform a large amount of symbolic processing and logical inferencing (as is being done in AI) does not result in intelligent behaviour. The main difference between human and machine intelligence comes from the fact that humans perceive everything as a *pattern*, whereas for a machine all are *data*. Even in routine data consisting of integer numbers (like telephone numbers, bank account numbers, car numbers), humans tend to see a pattern. Recalling the data is also normally from a stored pattern. If there is no pattern, then it is very difficult for a human being to remember and reproduce the data later. Thus storage and recall operations in humans and machines are performed by different mechanisms. The pattern nature in storage and recall automatically gives robustness and fault tolerance for a human system. Moreover, typically far fewer patterns than the estimated capacity of human memory systems are stored.

Functionally also humans and machines differ in the sense that humans *understand* patterns, whereas machines can be said to *recognize* patterns in data. In other words, humans can get the whole object in the data even though there may be no clear identification of subpatterns in the data. For example, consider the name of a person written in a handwritten cursive script. Even though individual patterns for each letter may not be evident, the name is understood due to the visual hints provided in the written script. Likewise, speech is understood even though the patterns corresponding to individual sounds may be distorted sometimes to unrecognizable extents. Another major characteristic of a human being is the ability to continuously learn from examples, which is not well understood at all in order to implement it in an algorithmic fashion in a machine.

Human beings are capable of making mental patterns in their biological neural network from input data given in the form of numbers, text, pictures, sounds etc., using their sensory mechanisms of vision, sound, touch, smell and taste. These mental patterns are formed even when the data are noisy, or deformed due to variations such as translation, rotation and scaling. The patterns are also formed from a temporal sequence of data as in the case of speech and motion pictures. Humans have the ability to recall the stored patterns even when the input information is noisy or partial (incomplete) or mixed with information pertaining to other patterns.

## 2.3 Pattern recognition tasks

The inherent differences in information handling by human beings and machines in the form of patterns and data, and in their functions in the form of understanding and recognition have led us to identify and discuss several pattern recognition tasks which human beings are able to perform very naturally and effortlessly, whereas we have no simple algorithms to implement these tasks on a machine. The identification of these tasks below is somewhat influenced by the organization of the artificial neural network models which we will be describing later in this paper.

**2.3a Pattern association:** Pattern association problem involves storing a set of patterns or a set of input-output pattern pairs in such a way that when test data are presented, the pattern or pattern pair corresponding to the data is recalled. This is purely a memory function to be performed for patterns and pattern pairs. Typically,

it is desirable to recall the correct pattern even though the test data are noisy or incomplete. The problem of storage and recall of patterns is called autoassociation. Since this is a content addressable memory function, the system should display *accretive* behaviour, i.e., should recall the stored pattern closest to the given input. It is also necessary to store as many patterns or pattern pairs as possible in a given system.

Printed characters or any set of fixed symbols could be considered as examples of patterns for these tasks. Note that the test patterns are the same as the training patterns, but with some noise added, or some portions missing. In other words, the test data are generated from the same source in an identical manner as the training data.

**2.3b Pattern mapping:** In pattern mapping, given a set of input patterns and the corresponding output pattern or class label, the objective is to capture the implicit relationship between the patterns and the output, so that when a test input is given, the corresponding output pattern or the class label is retrieved. Note that the system should perform some kind of *generalization* as opposed to *memorizing* the information. This can also be viewed as a pattern classification problem belonging to *supervised* learning category. Typically, in this case the test patterns belonging to a class are not the same as the training patterns, although they may originate from the same source. Speech spectra of steady vowels generated by a person, or hand-printed characters, could be considered as examples of patterns for pattern mapping problems. Pattern mapping generally displays *interpolative* behaviour, whereas pattern classification displays *accretive* behaviour.

**2.3c Pattern grouping:** In this case, given a set of patterns, the problem is to identify the subset of patterns possessing similar distinct features and group them together. Since the number of groups and the features of each group are not explicitly stated, this problem belongs to the category of *unsupervised* learning or pattern clustering. Note that this is possible only when the features are unambiguous as in the case of hand-printed characters or steady vowels. In the pattern mapping problem the patterns for each group are given separately, and the implicit, although distinct, features have to be captured through the mapping. In pattern grouping on the other hand, patterns belonging to several groups are given, and the system has to resolve the groups.

Examples of the patterns for this task could be printed characters or hand-printed characters. In the former case, the grouping can be made based on the data themselves. Moreover, in that case the test data are also generated from an identical source as the training data. For hand-printed characters or steady vowel patterns, the features of the patterns in the data are used for grouping. Therefore in this case the test data are generated from a similar source as the training data, so that only features are preserved and not necessarily the actual data values.

**2.3d Feature mapping:** In several patterns the features are not unambiguous. In fact the features vary over a continuum, and hence it is difficult to form groups of patterns having some distinct features. In such cases, it is desirable to display the feature changes in the patterns directly. This again belongs to the unsupervised learning category. In this case what is learnt is the *feature map* of a pattern and not the group or class to which the pattern may belong. This occurs, for example, in the

speech spectra for vowels in continuous speech. Due to changes in the vocal tract shape for the same vowel occurring in different contexts, the features (formants or resonances of the vocal tract in this case) vary over overlapping regions for different vowels.

2.3e *Pattern variability*: There are many situations when the features in the pattern undergo unspecified distortions each time the pattern is generated by the system. This can be easily seen in the normal handwritten cursive script. Human beings are able to recognize them due to some implicit interrelations among the features, which themselves cannot be articulated precisely. Classification of such patterns falls into the category of pattern variability task.

2.3f *Temporal patterns*: All the tasks discussed so far refer to the features present in a given static pattern. Human beings are able to capture effortlessly the dynamic features present in a sequence of patterns. This is true, for example, in speech where the changes in the resonance characteristics of the vocal tract system (e.g. formant contours) capture the significant information about the speech message. This is also true in any dynamic scene situation. All such situations require handling sequences of static patterns simultaneously, looking for changes in the features in the subpatterns in adjacent pattern pairs.

2.3g *Stability-plasticity dilemma*: In any pattern recognition task the input patterns keep changing. Therefore it is difficult to freeze the categorization task based on a set of patterns used in the training set. If it is frozen, then the system cannot learn the category that a new pattern may suggest. In other words, the system lacks its *plasticity*. On the other hand, if the system is allowed to change its categorization continuously, based on new input patterns, it cannot be used for any application such as pattern classification or clustering, as it is not *stable*. This is called *stability-plasticity dilemma* in pattern recognition.

## 2.4 *Methods for pattern recognition tasks*

Methods for solving pattern recognition tasks generally assume a sequential model for the pattern recognition process, consisting of pattern environment, sensors to collect data from the environment, feature extraction from the data and association/storage/classification/clustering using the features.

The simplest solution to a pattern recognition problem is to use template matching, where the data of the test pattern are matched point by point with the corresponding data in the reference pattern. Obviously, this can work only for very simple and highly restricted pattern recognition tasks. At the next level of complexity, one can assume a deterministic model for the pattern generation process, and derive the parameters of the model from given data in order to represent the pattern information in the data. Matching test and reference patterns are done at the parametric level. This works well when the model of the generation process is known with reasonable accuracy. One could also assume a stochastic model for the pattern generation process, and derive the parameters of the model from a large set of training patterns. Matching between test and reference patterns can be performed by several statistical methods like likelihood ratio, variance weighted distance, Bayesian classification etc. Other approaches for pattern recognition tasks depend on extracting features from



parameters or data. These features may be specific for the task. A pattern is described in terms of features, and pattern matching is done using descriptions of the features. Another method based on descriptions is called syntactic or structural pattern recognition in which a pattern is expressed in terms of primitives suitable for the classes of pattern under study (Schalkoff 1992). Pattern matching is performed by matching the descriptions of the patterns in terms of the primitives. More recently, methods based on the knowledge of the sources generating the patterns are being explored for pattern recognition tasks. These knowledge-based systems express knowledge in the form of rules for generating and perceiving patterns.

The main difficulty in each of the pattern recognition techniques alluded to above is that of choosing an appropriate model for the pattern generating process and estimating the parameters of the model in the case of a model-based approach, or extraction of features from data/parameters in the case of feature-based methods, or selecting appropriate primitives in the case of syntactic pattern recognition, or deriving rules in the case of a knowledge-based approach. It is all the more difficult when the test patterns are noisy and distorted versions of the patterns used in the training process. The ultimate goal is to impart to a machine the pattern recognition capabilities comparable to those of human beings. This goal is difficult to achieve using most of the conventional methods, because, as mentioned earlier, these methods assume a sequential model for the pattern recognition process. On the other hand, the human pattern recognition process is an integrated process involving the use of biological neural processing even from the stage of sensing the environment. Thus the neural processing takes place directly on the data for feature extraction and pattern matching. Moreover, the large size (in terms of number of neurons and interconnections) of the biological neural network and the inherently different mechanism of processing are attributed to our abilities of pattern recognition in spite of variability and noise in the data. Moreover, we are able to deal effortlessly with temporal patterns and also with the so-called stability-plasticity dilemma as well.

It is for these reasons attempts are being made to explore new models of computing, inspired by the structure and function of the biological neural network. Such models for computing are based on artificial neural networks, the basics of which are introduced in the next section.

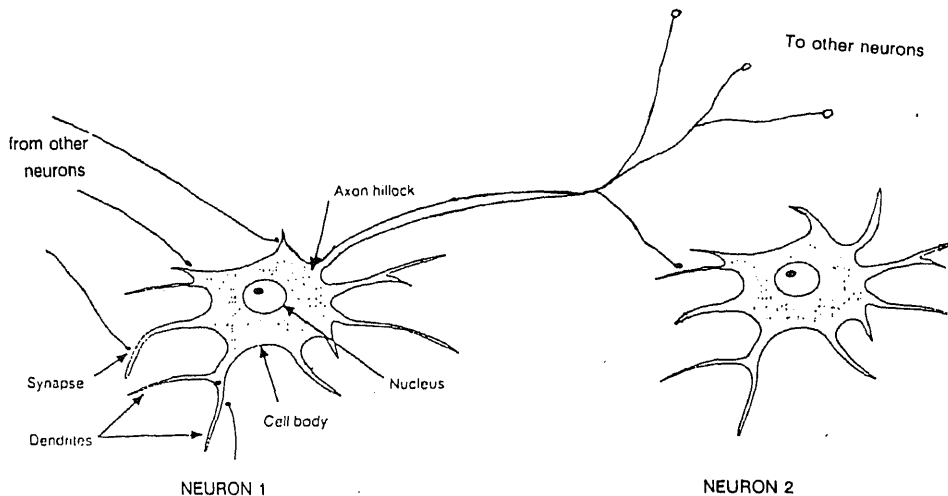
### 3. Basics of artificial neural networks

#### 3.1 Characteristics of biological neural networks

New models of computing to perform pattern recognition tasks based on our biological neural network are not expected to reach anywhere near the performance of the biological network for several reasons. Firstly, we do not fully understand the operation of a biological neuron and the dynamics of the neural interconnections. Secondly, it is nearly impossible to simulate (i) the number of neurons and their interconnections as it exists in a biological network, and (ii) the dynamics of the network that determines the operation of the network.

The features that make the performance of a biological network superior to even the most sophisticated AI computer system for pattern recognition tasks are the following (Hertz *et al* 1991).

(a) Robustness and fault tolerance – The decay of nerve cells does not seem to affect the performance of the network significantly.



**Figure 1.** Schematic drawing of a typical neuron or nerve cell. It includes dendrites, the cell body and a single axon. Synapses connect the axons of neurons to various parts of other neurons.

- (b) **Flexibility** – The network automatically adjusts to a new environment without using any preprogrammed instruction set.
- (c) **Ability to deal with a variety of data situations** – The network can deal with information that is fuzzy, probabilistic, noisy or inconsistent.
- (d) **Collective computation** – The network can routinely perform many operations in parallel and also a given task in a distributed manner.

These features are attributed to the structure and function of a biological neural network (Muller & Reinhardt 1990). The fundamental unit of the network is called a neuron or nerve cell. Figure 1 shows a schematic of the structure of a neuron. It consists of a cell body or soma where the cell nucleus is located. Tree-like networks of nerve fibres, called dendrites, are connected to the cell body. Extending from the cell body is a single long fibre, called the axon, which eventually branches into strands and substrands, connecting to many other neurons at the synaptic junctions or synapses. The receiving end of these junctions on other cells can be found both on the dendrites and on the cell bodies themselves. The axon of a typical neuron makes a few thousand synapses with other neurons.

The transmission of a signal from one cell to another at a synapse is a complex chemical process, in which specific transmitter substances are released from the sending side of the junction. The effect is to raise or lower the electrical potential inside the body of the receiving cell. If this potential reaches a threshold, electrical activity in the form of short pulses takes place. When this happens, the cell is said to have fired. This electrical activity of fixed strength and duration is sent down the axon.

The dendrites serve as receptors for signals from adjacent neurons, whereas the axon's purpose is the transmission of the generated neural activity to other nerve cells or to muscle fibres. In the first case the term interneuron may be used, whereas the neuron in the latter case is called motor neuron. A third type of neuron, which receives information from muscles or sensory organs, such as the eye or ear, is called a receptor neuron.

Although all neurons operate on the same basic principle, there exist several different types of neurons, distinguished by the size and degree of branching of their dendritic trees, the length of their axons, and other structural details. The complexity of the human central nervous system is due to the vast number of neurons and their mutual connections. Connectivity is characterized by the complementary properties of convergence and divergence. In the human cortex every neuron is estimated to receive converging input on the average from about  $10^4$  synapses. On the other hand, each cell feeds its output into many hundreds of other neurons. The total number of neurons in the human cortex is estimated to be in the vicinity of  $10^{11}$ , and are distributed in layers over a full depth of cortical tissue at a constant density of about 150,000 neurons per square millimetre. Combined with the average number of synapses per neuron, this yields a total of about  $10^{15}$  synaptic connections in the human brain, the majority of which develop during the first few months after birth. The study of the properties of complex systems built of simple, identical units, may lead to an understanding of the mode of operation of the brain in its various functions, although we are still far from it.

The simplified schematic and uniform connectionist units offer a surprisingly rich structure when assembled in a closely interconnected network. We shall call such a network an *artificial neural network*. Since artificial neural networks are implemented on computers, it is worth comparing the processing capabilities of computers with that of the biological neural networks (Simpson 1990).

Neural networks are slow in processing information. The cycle time corresponding to execution of one step of a program in a computer is in the range of a few nanoseconds, whereas the cycle time corresponding to a neural event prompted by an external stimulus, is in the milliseconds range. Thus computers process information a million times faster.

Neural networks perform massively parallel operations. Most programs operate in a serial mode, one instruction after another, in a conventional computer, whereas the brain operates with massively parallel programs that have comparatively fewer steps.

Neural networks have large numbers of computing elements, and the computing is not restricted to within neurons. The conventional computer typically has one central processing unit where all the computing takes place.

Neural networks store information in the strengths of the interconnections. In a computer, information is stored in the memory which is addressed by its location. New information is added by adjusting the interconnection strengths without completely destroying the old information, whereas in a computer the information is strictly replaceable.

Neural networks distribute the encoded information throughout the network, and hence they exhibit fault tolerance. In contrast, computers are inherently not fault tolerant, in the sense that information corrupted in the memory cannot be retrieved.

There is no central control in processing information in the brain. Thus there is no specific control mechanism external to the computing task. In a computer, on the other hand, there is a control unit which monitors all the activities of computing.

While the superiority of the human information processing system over the conventional computer for pattern recognition tasks stems from the basic structure and operation of the biological neural network, it is possible to realize some of the features of the human system using an artificial neural network consisting of basic computing elements. In particular, it is possible to show that such a network exhibits

parallel and distributed processing capability. In addition, information can be stored in a distributed manner in the connection strengths so as to achieve fault tolerance.

### 3.2 Artificial neural networks – terminology

**3.2a Processing unit:** We can consider an artificial neural network (ANN) as a highly simplified model of the structure of the biological neural network. An ANN consists of interconnected *processing units*. The general model of a processing unit consists of a summing part followed by an output part. The summing part receives  $n$  input values, weighs each value, and performs a weighted sum. The weighted sum is called the *activation value*. The sign of the weight for each input determines whether the input is *excitatory* (positive weight) or *inhibitory* (negative weight). The inputs could be discrete or continuous data values, and likewise the outputs also could be discrete or continuous. The input and output may also be viewed as deterministic or stochastic or fuzzy, depending on the nature of the problem and its solution.

**3.2b Interconnections:** In an artificial neural network several processing units are interconnected according to some topology to accomplish a pattern recognition task. Therefore the inputs to a processing unit may come from outputs of other processing units, and/or from an external source. The output of each unit may be given to several units including itself. The amount of the output of one unit received by another unit depends on the strength of the connection between the units, and it is reflected in the *weight* value associated with the connecting link. If there are  $N$  units in a given ANN then at any instant of time each unit will have a unique activation value and a unique output value. The set of the  $N$  activation values of the network defines the *activation state* of the network at that instant. Likewise, the set of the  $N$  output values of the network define the *output state* of the network at that instant. Depending on the discrete or continuous nature of the activation and output values, the state of the network can be described by a point in a discrete or continuous  $N$ -dimensional space.

**3.2c Operations:** In operation, each unit of an ANN receives inputs from other connected units and/or from an external source. A weighted sum of the inputs is computed at a given instant of time. The resulting activation value determines the actual output from the *output function* unit, i.e., the output state of the unit. The output values and other external inputs in turn determine the activation and output states of the other units. The activation values of the units (activation state) of the network as a function of time are referred to as *activation dynamics*. The activation dynamics also determine the dynamics of the output state of the network. The set of all activation states defines the *state space* of the network. The set of all output states defines the *output* or *signal state space* of the network. Activation dynamics determines the trajectory of the path of the states in the state space of the network.

For a given network, defined by the units and their interconnections with appropriate weights, the activation states refer to the *short term memory* function of the network. Generally the activation dynamics is followed to *recall* a pattern stored in a network.

In order to store a pattern in a network, it is necessary to adjust the weights of the network. The sets of all weight values (corresponding to strengths of all connecting links of an ANN) defines the *weight space*. If the weights are changing, then the set of

weight values as a function of time defines the *synaptic dynamics* of the network. Synaptic dynamics is followed to adjust the weights in order to store given patterns in the network. The process of adjusting the weights is referred to as *learning*. Once the learning process is completed, the final set of weight values corresponds to the *long term memory* function of the network. The procedure to incrementally update each of the weights is called a *learning law* or *learning algorithm*.

**3.2d Update:** In implementation, there are several options available for both activation and synaptic dynamics. In particular, the updating of the output states of all units could be performed *synchronously*. In this case, the activation values of all units are computed at the same time assuming a given output state throughout. From these activation values the new output state of the network is derived. In an *asynchronous* update, on the other hand, each unit is updated sequentially, taking the current output state of the network into account each time. For each unit, the output state can be determined from the activation value either *deterministically* or *stochastically*.

In practice, the activation dynamics, including the update, is much more complex in a biological neural network. The ANN models along with the equations governing the activation and synaptic dynamics are developed according to the complexity of the pattern recognition task to be handled.

### 3.3 Models of neurons

In this section we will consider three classical models for an artificial neuron or processing unit.

**3.3a McCulloch–Pitts model:** In the McCulloch–Pitts (MP) model (figure 2) the activation ( $x$ ) is given by a weighted sum of its  $n$ -input signal values  $\{a_i\}$  and a bias term ( $\theta$ ). The activation could have an additional absolute inhibition term, which can prevent excitation of the neuron. The output signal ( $s$ ) is typically a nonlinear function of the activation value. Three common nonlinear functions (binary, ramp and sigmoid) are shown in figure 3, although the binary function was used in the original MP model. The following equations describe the operation of an MP model:

$$\text{activation:} \quad x = \sum_{i=1}^n w_i a_i - \theta - [\text{inhibition}],$$

$$\text{output signal:} \quad s = f(x).$$

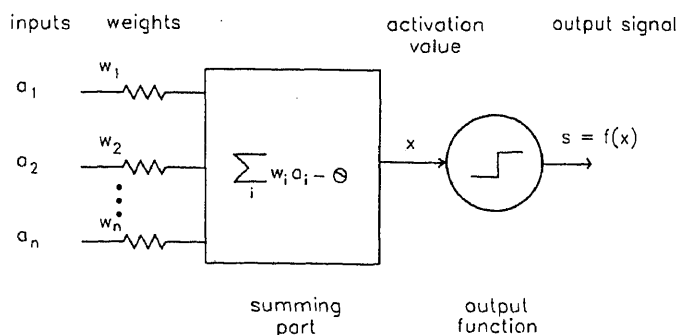


Figure 2. The McCulloch–Pitts model of a neuron.

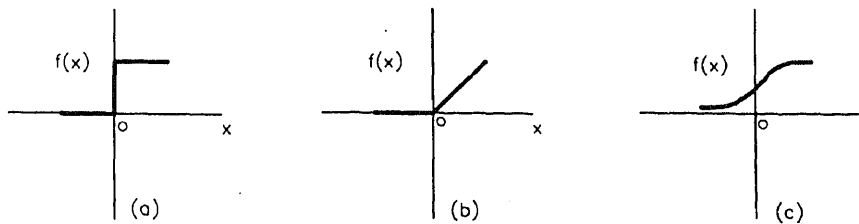


Figure 3. Some nonlinear functions. (a) Binary, (b) ramp and (c) sigmoid.

In this model the weights  $w_i$  are constant. That means there is no learning. Networks consisting of MP neurons with binary (on-off) output signals can be configured to perform several logical functions (McCulloch & Pitts 1943).

3.3b *Perceptron*: Rosenblatt's perceptron model (figure 4) for an artificial neuron consists of outputs from sensory units to a fixed set of association units, the outputs of which are fed to an MP neuron (Rosenblatt 1958). The association units perform predetermined manipulations on their inputs. The main deviation from the MP model is that here learning (i.e., adjustment of weights) is incorporated in the operation of the unit. The target output ( $b$ ) is compared with the actual binary output ( $s$ ) and the error is used to adjust the weights (Rosenblatt 1962). The following equations describe the operation of the perceptron model of a neuron.

$$\text{activation:} \quad x = \sum_{i=1}^n w_i a_i - \theta,$$

$$\text{output signal:} \quad s = f(x),$$

$$\text{error:} \quad \delta = b - s,$$

$$\text{weight update:} \quad \frac{dw_i}{dt} = \eta \delta a_i,$$

where  $\eta$  is called learning rate parameter.

There is the perceptron learning law which gives a step-by-step procedure for adjusting the weights. Whether the adjustment converges or not depends on the

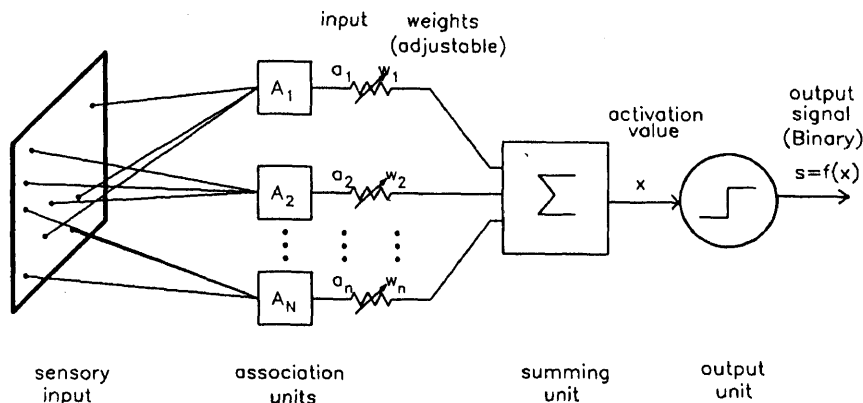
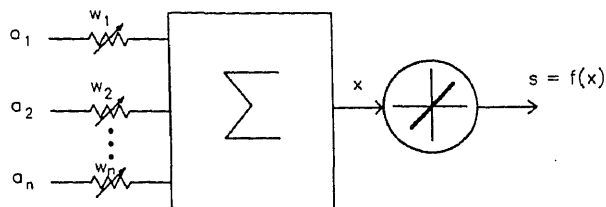


Figure 4. Rosenblatt's model of a neuron.



**Figure 5.** Widrow's adaline model of a neuron.

nature of the desired input–output pair to be represented by the model. The perceptron convergence theorem (Rosenblatt 1962) enables us to determine whether a given pattern pair is representable or not. If the weight values converge, then the corresponding problem is said to be *representable* by the perceptron network.

**3.3c Adaline:** The main distinction between Rosenblatt's perceptron model and Widrow's adaline model (figure 5) is that in the adaline model the analog activation value ( $x$ ) is compared with the target output ( $b$ ). In other words, the output is a linear function of the activation value ( $x$ ). The equations that describe the operation of an adaline are as follows (Widrow & Hoff 1960):

$$\text{activation:} \quad x = \sum_{i=1}^n w_i a_i - \theta,$$

$$\text{output signal:} \quad s = f(x) = x,$$

$$\text{error:} \quad \delta = b - s = b - x,$$

$$\text{weight update:} \quad dw_i/dt = \eta \delta a_i.$$

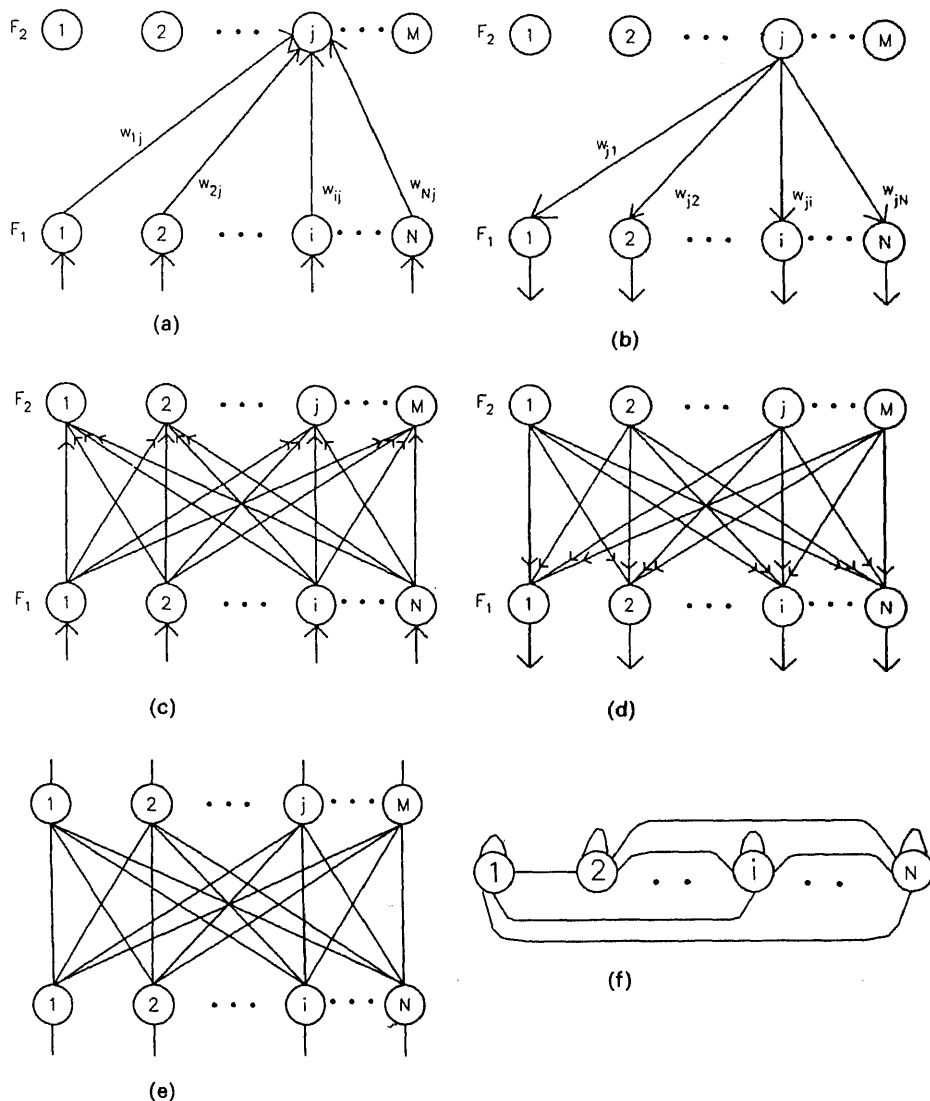
This rule minimizes the mean squared error  $\delta^2$ , averaged over all inputs. Hence it is called the *least mean squared* (LMS) error learning law. The law is derived using the negative gradient of the error surface in the weight space. Hence it is also called a *gradient descent algorithm*.

### 3.4 Topology

Artificial neural networks are useful only when the processing units are organized in a suitable manner to accomplish a given pattern recognition task. This section presents a few basic structures which will assist in evolving new architectures. The arrangement of the processing units, connections, and pattern input/output is referred to as *topology* (Simpson 1992, pp. 3–24).

Artificial neural networks are normally organized into layers of processing units. Connections can be made either from units of one layer to units of another (interlayer connections) or from the units within the layer (intralayer connections) or both inter and intralayer connections. Further, the connections among the layers and among the units within a layer can be organized either in a feedforward manner or in a feedback manner. In a feedback network the same processing unit may be visited more than once.

We will discuss a few basic structures which form building blocks for complex neural network architectures. Let us consider two layers  $F_1$  and  $F_2$  with  $N$  and  $M$  processing units, respectively. By providing connections to the  $j$ th unit in  $F_2$  from all the units in  $F_1$ , as shown in figures 6a and b, we get two network structures instar and outstar, which have fan-in and fan-out geometries, respectively. The units in the



**Figure 6.** Some basic structures of the Artificial Neural Networks. (a) Instar, (b) outstar, (c) group of instars, (d) group of outstars, (e) bidirectional associative memory, and (f) autoassociative memory.

$F_1$  layer are linear units, so that for each unit  $i$  in this layer the input ( $a_i$ ) = activation ( $x_i$ ) = output signal ( $s_i$ ). In instar, during learning, the weight vector  $\mathbf{w}_j(w_{j1}, w_{j2}, \dots, w_{jN})$  is adjusted so as to approach the given input vector  $\mathbf{a}$  at  $F_1$  layer. Therefore whenever the input is given to  $F_1$ , then the  $j$ th unit of  $F_2$  will be activated to the maximum extent. Thus the operation of the instar can be viewed as content addressing the memory. In the case of the outstar, during learning, the weight vector for the connections from the  $j$ th unit in  $F_2$  approaches the activity pattern in  $F_1$  when input vector  $\mathbf{a}$  is present at  $F_1$ . During recall, whenever the unit  $j$  is activated, the signal pattern ( $s_j w_{j1}, s_j w_{j2}, \dots, s_j w_{jN}$ ) will be transmitted to  $F_1$ , which then produces the original activity pattern corresponding to the input vector  $\mathbf{a}$ , although the input is absent. Thus the operation of the outstar can be viewed as memory addressing the contents.



When all the connections from units in  $F_1$  and  $F_2$  are made as in figure 6c, then we obtain a heteroassociation network. This network can be viewed as a group of instars, if the flow is from  $F_1$  to  $F_2$ . On the other hand, if the flow is from  $F_2$  to  $F_1$ , then the network can be viewed as a group of outstars (figure 6d).

When the flow is bidirectional, and the weights are symmetric  $w_{ij} = w_{ji}$ , then we get a bidirectional associative memory (figure 6e), where either of the layers can be used as input/output.

If the two layers  $F_1$  and  $F_2$  coincide, then we obtain an autoassociative memory in which each unit is connected to every other unit and to itself (figure 6f).

### 3.5 Activation and synaptic dynamics

Artificial neural networks can be considered as trainable nonlinear dynamical systems (Kosko 1972). For a network consisting of  $N$  processing units, the activation state of the network at any given instant corresponds to a point in the  $N$ -dimensional state space. The dynamics of the neural network traces a trajectory in the state space. The trajectory begins with a point in the state space representing a computational problem and ends at a point in the state space representing a computational solution. Most of the trajectory corresponds to the transient behaviour of computations. The trajectory ends at an *equilibrium state* of the system in the normal course. An equilibrium state is one at which small perturbations around it due to neuronal dynamics will not perturb the state.

Neuronal dynamics consists of two parts: one corresponding to the dynamics of activation states and the other corresponding to the dynamics of synaptic weights. The *activation dynamics* determines the time evolution of the neuronal activations, and it is described by a system of first order differential equations. The equations governing the dynamics are described in terms of the first derivative of the activation state, i.e.,  $dx_i/dt$ . Likewise synaptic dynamics determines the changes in the synaptic weights. The equations governing the dynamics are described in terms of the first derivative of the synaptic weights, i.e.,  $dw_{ij}/dt$ , where  $w_{ij}$  is the strength of the connecting link from the  $j$ th unit to the  $i$ th unit. Synaptic weights change gradually, whereas the neuronal activations fluctuate rapidly. Therefore, while computing the activation dynamics, the synaptic weights are assumed to be constant. The synaptic dynamics dictates the *learning process*. The short term memory (STM) in neural networks is modelled by the activation state of the network. The long term memory (LTM) corresponds to the encoded pattern information in the synaptic weights due to learning.

**3.5a Models of activation dynamics:** Different models are proposed for the activation dynamics, the most common ones among them are the additive and shunting activation models. The additive activation model is given by the equation for the rate of change of the activation of the  $i$ th unit as (Grossberg 1988; Carpenter 1989).

$$dx_i/dt = -x_i + \sum[\text{excitatory inputs}] - \sum[\text{inhibitory inputs}].$$

In this equation the first term on the right hand side contributes to a passive decay term. The net excitatory and inhibitory inputs are contributed by signals from other units appropriately weighted by the synaptic strengths and by the externally applied inputs.

In the steady state there will not be any change in activation. That is  $dx_i/dt = 0$ . In such a case the activation value is given by the net excitatory and inhibitory inputs. That is

$$x_i = \sum [\text{excitatory inputs}] - \sum [\text{inhibitory inputs}].$$

For a specific case  $x_i$  can be written as

$$x_i = \sum_j w_{ij}s_j - \theta_i + I_i.$$

The sign of  $w_{ij}$  determines whether the contribution is excitatory or inhibitory.  $\theta_i$  is a fixed bias term for the unit, and it becomes the resting value in the absence of all inputs.  $I_i$  is the net external input to the unit  $i$ . The sign of  $I_i$  determines whether it is excitatory or inhibitory.

An important generalization of the additive model is the shunting activation model given by the equation (Grossberg 1988),

$$dx_i/dt = -x_i + (A - x_i)\sum [\text{excitatory inputs}] - (B + x_i)\sum [\text{inhibitory inputs}],$$

where the activity  $x_i$  remains bounded in the range  $(-B, A)$ , and it decays to the resting level 0 in the absence of all inputs. In this model the excitatory inputs drive the activity towards a finite maximum  $A$ , and the inhibitory inputs drive the activity towards a finite minimum  $-B$ . The shunting model represents a special case of Hodgkin-Huxley membrane equations to describe the physiology of single nerve cell dynamics (Hodgkin & Huxley 1952).

The activation models considered so far are called *deterministic models*. In practice, the input/output patterns and the activation values can be considered as samples of a random process, and the output signal of each unit may be a random function of the unit's activation value. In such a case the network activation state can be viewed as a *vector stochastic process*. Each unit in turn behaves as a *scalar stochastic process* (Kosko 1992).

**3.5b Models of synaptic dynamics:** Synaptic dynamics is described in terms of expressions for the first derivative of the weights. They are called *learning equations* (Kosko 1992). Typical (basic) learning involves adjustment of the weight vector such that

$$\Delta \mathbf{w}_i(t) = \eta g[\mathbf{w}_i(t), \mathbf{a}(t), b_i(t)] \mathbf{a}(t),$$

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \Delta \mathbf{w}_i(t),$$

where

$\eta$  = learning rate parameter,

$\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{iN}]^T$  weight vector with components  $w_{ij}$ ,

$w_{ij}$  = weight connecting the  $j$ th input unit to the  $i$ th processing unit,

$\mathbf{a}$  = input vector with components  $a_i, i = 1, 2, \dots, N$ ,

$\mathbf{b}$  = desired output vector with components  $b_i, i = 1, 2, \dots, M$ .

Input units are assumed linear. Hence  $\mathbf{a} = \mathbf{x}$  (unit activation) =  $\mathbf{s}$  (unit output).

Output units are in general nonlinear. Hence  $s_i = f(\mathbf{w}_i^T \mathbf{a})$ .

The function  $g$  may be viewed as a learning function that depends on the type of learning adopted.

Continuous time learning can be expressed

$$\frac{d\mathbf{w}_i(t)}{dt} = \eta g[\mathbf{w}_i(t), \mathbf{a}(t), b_i(t)] \mathbf{a}(t).$$

In discrete time learning, at the  $k$ th step the new weight is given by

$$\mathbf{w}_i^{k+1} = \mathbf{w}_i^k + \eta g[\mathbf{w}_i^k, \mathbf{a}^k, b_i^k] \mathbf{a}^k.$$

There are different methods for implementing the synaptic dynamics. These methods are called learning laws. A few common discrete time learning laws are given below (Zurada 1992).

### 3.5c Hebb's law (Hebb 1949):

Here  $g(.) = f(\mathbf{w}_i^T \mathbf{a})$ , where  $f$  is the output function. Therefore

$$\begin{aligned} \Delta w_{ij} &= \eta f'(\mathbf{w}_i^T \mathbf{a}) a_j \\ &= \eta s_i a_j, \quad \text{for } j = 1, 2, \dots, N \end{aligned}$$

This law requires weight initialization  $\mathbf{w}_i \approx \mathbf{0}$  prior to learning.

### 3.5d Perceptron learning law (Rosenblatt 1962):

Here  $g(.) = b_i - s_i = b_i - \text{sgn}(\mathbf{w}^T \mathbf{a})$ . Therefore

$$\Delta w_{ij} = \eta [b_i - \text{sgn}(\mathbf{w}^T \mathbf{a})] a_j, \quad \text{for } j = 1, 2, \dots, N.$$

This rule is applicable for bipolar output function. The weights can be initialized to any values prior to learning.

### 3.5e Delta learning law:

Here  $g(.) = [b_i - f(\mathbf{w}_i^T \mathbf{a})] f'(\mathbf{w}_i^T \mathbf{a})$ . This is obtained by setting

$$\Delta \mathbf{w}_i = -\eta \nabla E$$

where  $-\nabla E$  is the negative gradient of the error  $E = \frac{1}{2} [b_i - f(\mathbf{w}_i^T \mathbf{a})]^2$ . Therefore

$$\Delta w_{ij} = \eta (b_i - s_i) f'(\mathbf{w}_i^T \mathbf{a}) a_j, \quad \text{for } j = 1, 2, \dots, N$$

Here  $f(.)$  is a continuous function. The weights may be initialized to any values.

### 3.5f Widrow-Hoff LMS learning law (Widrow & Hoff 1960):

Here  $g(.) = b_i - \mathbf{w}_i^T \mathbf{a}$ . Therefore

$$\Delta w_{ij} = \eta (b_i - \mathbf{w}_i^T \mathbf{a}) a_j, \quad \text{for } j = 1, 2, \dots, N.$$

This is a special case of the delta learning law where the output function is assumed to be linear, i.e.,  $f(\mathbf{w}_i^T \mathbf{a}) = \mathbf{w}_i^T \mathbf{a}$ . The weights may be initialized to any values.

## 3.5g Correlation learning law:

$$\Delta w_{ij} = \eta b_i a_j, \quad \text{for } j = 1, 2, \dots, N.$$

This is applicable for binary output units. This is a special case of Hebbian learning with output signal  $(s_i) = \text{desired signal } (b_i)$ . The weights are initialized to zero prior to learning.

## 3.5h Instar (winner-take-all) learning law (Grossberg 1982):

$$\Delta w_{mj} = \eta (a_j - w_{mj}), \quad \text{for } j = 1, 2, \dots, N,$$

where  $w_m^T \mathbf{a} = \max_i (w_i^T \mathbf{a})$ . Here the weights are initialized to random values prior to learning and their lengths are normalized during learning.

## 3.5i Outstar learning law (Grossberg 1982):

$$\Delta w_{kj} = \eta (b_k - w_{kj}), \quad \text{for } k = 1, 2, \dots, K$$

where  $\mathbf{b}$  is the desired response from the layer of  $K$  neurons. The weights are initialized to zero before learning.

There are several learning laws in use, and new laws are being developed to suit a given application and architecture. Some of these will be discussed in the appropriate sections later. But there are some general categories that these laws fall into, based on the characteristics they are expected to possess for different applications. In first place, the learning or weight changes could be *supervised* or *unsupervised*. In supervised learning the weight changes are determined by the difference between the desired output and the actual output. Some of the supervised learning laws are: error correction learning or delta rule, stochastic learning, and hardwired systems (Simpson 1992, pp. 3–24). Supervised learning may be used for *structural learning* or for *temporal learning*. Structural learning is concerned with capturing in the weights the relationship between a given input-output pattern pair. Temporal learning is concerned with capturing in the weights the relationship between neighbouring patterns in a sequence of patterns.

Unsupervised learning discovers features in a given set of patterns and organizes the patterns accordingly. There is no externally specified desired output as in the case of supervised learning. Examples of unsupervised learning laws are: Hebbian learning, differential Hebbian learning, principle component learning and competitive learning (Simpson 1992, pp. 3–24). Unsupervised learning uses mostly local information to update the weights. The local information consists of signal or activation values of the units at either end of the connection for which the weight update is being made.

Learning methods can be grouped into *off-line* and *on-line*. In off-line learning all the given patterns are used, may be several times if needed, to adjust the weights. Most error correction learning laws belong to the off-line category. In on-line learning each new pattern or set of patterns can be incorporated into the network without any loss of the prior stored information. Thus an on-line learning allows the neural network to add new information continuously. An off-line learning provides superior solutions because information is extracted when all the training patterns are

available, whereas an on-line learning updates only the available information of the past patterns in the form of weights.

In practice, the training patterns can be considered as samples of random processes. Learning laws could take into account the changes in the random process reflected through the samples patterns. Thus one could define stochastic versions of the deterministic learning laws described so far. The random learning laws are expressed as first order stochastic differential equations. For example, the random signal Hebbian learning law relates random processes as (Kosko 1992)

$$dw_{ij}/dt = -w_{ij} + \eta s_i s_j + n_{ij},$$

where the output random process  $\{s_i\}$  is a result of the signal random process  $\{s_j\}$ , which in turn may be a result of another activation random process caused by the input process.  $\{n_{ij}\}$  can be assumed to be a zero-mean Gaussian white noise process.

In supervised learning one can derive a stochastic approximation to the learning law using the following argument: Given a set of  $L$  random samples, each sample consisting of the pattern pairs  $(\mathbf{a}_l, \mathbf{b}_l)$ , a supervised learning attempts to minimize an unknown error functional  $E[\delta_l]$ , where  $\delta_l$  is the error between the desired output and the actual output signal. The gradient of  $-E[\delta_l]$  points in the direction of steepest descent on the unknown expected error surface. Since the joint probability density function of the input/output pattern pairs is not known, only the error  $\delta_l$  is used as an estimate of  $E[\delta_l]$ . Since  $\delta_l$  is also a random process, for each iteration in a discrete stochastic gradient descent algorithm, the weight update at the  $(k+1)$ th iteration is given by (Kosko 1992)

$$w_{ij}^{k+1} = w_{ij}^k - \eta \delta_{li}^k a_{lj},$$

where  $\delta_{li}^k = b_{li} - s_i^k$ . Since the given data are sample functions of a random process, the corresponding weights at each iteration are also random.

*Synaptic equilibrium* in the deterministic signal Hebbian law occurs in the steady state when the weights stop changing. That is,

$$dw_{ij}/dt = 0, \quad \text{for all } i, j.$$

In the stochastic case the synaptic weights reach a *stochastic equilibrium* when the changes in the weights are contributed by only the random noise. That is, at stochastic equilibrium, the expectation or ensemble average of the change in weights is given by (Kosko 1990)

$$E[(dw_{ij}/dt)^2] = \sigma_{ij}^2,$$

where  $\sigma_{ij}^2$  is the variance of the noise process  $n_{ij}$ .

**3.5j Stability and convergence:** So far the activation and synaptic dynamics equations are described in terms of first-order differential equations which are continuous time equations. Discrete time versions of these equations are convenient for implementation of the network dynamics on a digital computer. In discrete time implementation the activation state of each unit at each stage is computed in terms of the state of the network in the previous stage. The state update at each stage could be made asynchronously, i.e. each unit is updated using the new updated state, or synchronously, i.e., all the units are updated using the same previous state.

The implications of these implementations are on the *stability* of the equilibrium activation states of a feedback neural network, and on the *convergence* of the synaptic weights while minimizing the error between the desired output and the actual output during learning. In general, there are no standard methods to determine whether network activation dynamics or synaptic dynamics leads to stability or convergence, respectively, or not (Kosko 1992; Simpson 1992, pp. 3–24).

**3.5k Neural network recall:** During learning, the weights are adjusted to store the information in a given pattern or a pattern pair. However, during performance, the weight changes are suppressed, and the input to the network determines the output activation  $x_j$  or signal values  $s_j$ . This operation is called *recall* of stored information. The recall techniques are different for feedforward and feedback networks.

The simplest feedforward network uses the following equation to compute the output signal from the input data vector  $\mathbf{a}$  to the input layer  $F_1$ :

$$s_i = f_i \left( \sum_j w_{ij} a_j \right),$$

where  $f_i$  is the output function of the  $i$ th unit in the output layer  $F_2$ . Here the units in the input layer  $F_1$  are assumed to be linear.

A recall equation for a network with feedback connections is given by (Simpson 1992, pp. 3–24)

$$x_i(t+1) = (1 - \alpha)x_i(t) + \beta \sum_{j=1}^N f_j(x_j(t))w_{ij} + a_i,$$

where  $x_i(t+1)$  is the activation value of the  $i$ th unit in a single layer neural network at time  $(t+1)$ ,  $f_j$  is the nonlinear output function of the  $j$ th unit,  $\alpha$  is a positive constant that regulates the amount of decay the unit has during the update interval,  $\beta$  is a positive constant that regulates the amount of feedback the other units provide to the  $i$ th unit, and  $a_i$  is the external input to the  $i$ th unit. In general, stability is the main issue in feedback networks. If the network reaches a stable state in a finite number of iterations, then the resulting output signals represent the nearest neighbour stored pattern of the system for the approximate input pattern  $\mathbf{a}$ .

Cohen & Grossberg (1983) showed that for a wide class of neural networks with certain constraints, the network with fixed weights reaches a stable state in a finite period of time for any initial condition. Later Kosko showed that a neural network could learn and recall at the same time, and yet remain stable (Kosko 1990).

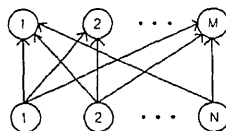
The response of a network due to recall could be the nearest neighbour or interpolative. In the nearest neighbour case, the stored pattern closest to the input pattern is recalled. This typically happens in the feedforward pattern classification or feedback pattern matching networks. In the interpolative case, the recalled pattern is a combination of the outputs corresponding to the input training patterns nearest to the given input test pattern. This happens in the feedforward pattern mapping networks.

#### 4. Functional units of ANN for pattern recognition tasks

So far we have considered issues in pattern recognition and introduced basics of artificial neural networks. In this section we discuss some functional units of artificial neural networks that are useful to solve simple pattern recognition tasks. In particular,

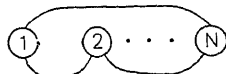
## 1. Feedforward ANN

- (a) Pattern association
- (b) Pattern classification
- (c) Pattern mapping/classification



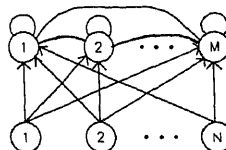
## 2. Feedback ANN

- (a) Autoassociation
- (b) Pattern storage (LTM)
- (c) Pattern environment storage (LTM)



## 3. Feedforward and Feedback ANN

- (a) Pattern storage (STM)
- (b) Pattern clustering
- (c) Feature map



**Figure 7.** Summary of ANN for pattern recognition problems.

we discuss artificial neural networks for the following pattern recognition problem and for various special cases of the problem.

## PROBLEM

Design a neural network to associate the pattern pairs  $(\mathbf{a}_1, \mathbf{b}_1)$ ,  $(\mathbf{a}_2, \mathbf{b}_2), \dots, (\mathbf{a}_L, \mathbf{b}_L)$ , where  $\mathbf{a}_l = (a_{l1}, a_{l2}, \dots, a_{lN})$  and  $\mathbf{b}_l = (b_{l1}, b_{l2}, \dots, b_{lM})$  are  $N$  and  $M$  dimensional vectors, respectively.

Figure 7 shows the organization of the networks and the pattern recognition tasks to be discussed in this section. We consider three types of ANN: Feedforward, feedback and a combination of both. We begin discussion of each network with only a minimal structure, and study their capabilities and limitations. To start with, the feedforward network consists of two layers of processing units, one layer with linear units for receiving the external input, and the other layer for delivering the output. A minimal feedback network consists of a set of processing units, each connected to all other units. A combination network consists of an input layer of linear units feeding to the output layer of units in a feedforward manner, and a feedback connection among the units in the output layer, including self feedback. We consider each one of these networks in some detail.

### 4.1 Pattern recognition tasks by feedforward ANN (figure 8)

**4.1a Pattern association:** The objective is to design a linear network that can capture the association in the pairs of vectors  $(\mathbf{a}_l, \mathbf{b}_l)$ ,  $l = 1, 2, \dots, L$ , through a set of weights to be determined by a learning or training law. The input data used in training are typically generated synthetically, like machine printed characters. The input data used for recall may be corrupted by external noise.

The network consists of a set of weights connecting the two layers of processing units, the output function of each unit being linear. Such a network is called a linear associator network. Due to linearity of the output function of each unit, the activation values and the output signals of the units in the input layer are same as the input

**Pattern association**

- \* *Arch*: Two layers, linear processing unit, single set of weights
- \* *Learning*: Hebb (orthogonal) rule, Delta (linearly independent) rule
- \* *Recall*: Direct
- \* *Limitation*: Linear independence, # patterns restricted to dimensionality
- \* *To overcome*: Nonlinear processing unit, becomes a pattern classification problem

**Pattern classification**

- \* *Arch*: Two layers, nonlinear processing units, geometrical interpretation
- \* *Learning*: Delta rule
- \* *Recall*: Direct
- \* *Limitation*: Linearly separable functions, hard problems
- \* *To overcome*: More layers, hard learning problems

**Pattern mapping/classification**

- \* *Arch*: Multilayer (hidden), nonlinear processing units, geometric interpretation
- \* *Learning*: Generalized delta rule – backpropagation
- \* *Recall*: Direct
- \* *Limitation*: Slow learning
- \* *To overcome*: More complex architectures

**Figure 8.** Pattern recognition tasks by feedforward ANN.

data values. The activation value of the  $i$ th unit in the output layer is given by

$$y_i = \sum_{j=1}^N w_{ij} a_{ij}.$$

The output of the  $i$ th unit is the same as its activation value  $y_i$ , since the output function of the unit is linear. The objective is to determine a set of weights  $w_{ij}$  in such a way that the actual output  $b'_{ii}$  is equal to the desired output  $b_{ii}$  for all the  $L$  pattern pairs.

If the input  $L$  pattern vectors  $\{\mathbf{a}_i\}$  are all orthogonal, then it is possible to use Hebb's learning law to determine the optimal weights of the network (Hecht-Nielsen 1990). Note that a learning law enables updating of weights as patterns are applied one by one to the network. The optimality of the weights is determined by minimizing the mean squared error between the desired and the actual output values. The optimal weights after  $l$  pattern pairs are fed to the network are given by

$$w_{ij}^l = w_{ij}^{l-1} + b_{ii} a_{ij}, \quad w_{ij}^0 = 0.$$

The final optimal weights for pattern association task are given by

$$w_{ij} = w_{ij}^L.$$

If the input vectors  $\{\mathbf{a}_i\}$  are only linearly independent, but not necessarily orthogonal, then the optimal weights that minimize the mean squared error can be obtained using the LMS learning law (Widrow & Hoff 1960; Hecht-Nielsen 1990).

Once the network is trained, for any given input pattern  $\mathbf{a}_i$ , the associated pattern  $\mathbf{b}_i$  can be recalled using the equations

$$y_i = \sum_{j=1}^N w_{ij} a_{ij} \quad \text{and} \quad b_{ii} = y_i.$$



When noisy input patterns are used during recall, i.e.,  $\{a_{ij}\}$ , then the recalled pattern  $\{b'_{ij}\}$  will also be noisy. Since the given set of input pattern  $\{a_i\}$ ,  $i = 1 \dots L$ , is assumed to be linearly independent, the number of patterns in the input set is limited to the dimensionality of the input vector, namely,  $N$ . Therefore, it is not possible to store more than  $N$  pattern pairs in a linear associative network. If the number of input pattern are more than its dimension ( $N$ ), or if the input set (even for  $L < N$ ) are not linearly independent, then the resulting weight vectors are not optimal any more. In such a case the recall of the associative pattern for a given input pattern may not be correct always.

Even if the input patterns are linearly independent and optimal weights are used, the recall may be in error if a noisy input pattern is presented to recall the associated pattern (Murakami & Aibara 1987).

In practice, linear independence is too severe a restriction to satisfy. Moreover the number of input patterns may far exceed the dimensionality of the input pattern space. It is possible to overcome these limitations by using *nonlinear* output functions in the processing units of the feedforward ANN. Once the restriction on the number of input patterns is removed, then the problem becomes a *pattern classification* problem, which we will discuss in the next section.

**4.1b Pattern classification:** In an  $N$ -dimensional space if a set of points could be considered as input patterns without restriction on their number, and if an output pattern, not necessarily distinct, is assigned to each of the input patterns, then the number of distinct output patterns can be viewed as distinct classes or class labels for the input patterns. Since there is no restriction on the type and number of input patterns, the input-output pattern pairs  $(a_i, b_i)$ ,  $i = 1, 2, \dots, L$  in this case can be considered as a training set for a pattern classification problem. Typically for pattern classification problems the output patterns are points in a discrete (normally binary)  $M$ -dimensional space. The input patterns are usually from natural sources like speech and hand-printed characters. The input patterns may be corrupted by external noise at the time of recall.

A two-layer network with nonlinear (threshold or hardlimiting) output function for the units in the output layer, can be used to perform the task of pattern classification. This may also be identified as a single layer perceptron network (Rosenblatt 1962). The network can be trained (i.e., weights can be adjusted) for the given set of input-output patterns using a delta rule.

The corresponding learning is also called *perceptron learning law* (Rosenblatt 1962; Minsky & Papert 1988). The training patterns are applied several times, if needed, until the weights do not change appreciably. But there is no guarantee that the weights will converge to some stable values. Convergence of the weights depends on whether the problem specified by the input-output pattern pairs is *representable* or not by a network of this type. For all representable problems the learning law converges.

During recall, a pattern generated from one of the same sources is given as input. By direct computation of the weighted sum of the input, the network determines the pattern class to which the input belongs. The network thus exhibits accretive behaviour. Even when the input pattern is noisy, the output class may still be correct, provided the noise has not significantly altered the input pattern.

The unrepresentable problems are called *hard problems*. Such problems arise if the function  $\phi$  relating the output and input ( $b_i = \phi(a_i)$ ) is not *linearly separable*. In

(a)

Input

 $a_1 \ a_2$ 

0 0

0 0

1 0

1 1

Output

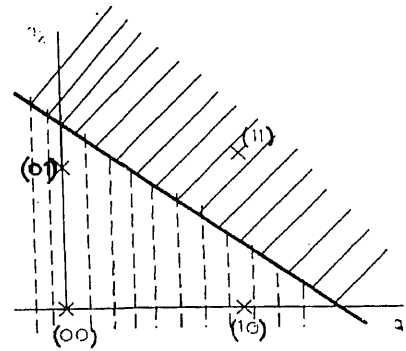
 $b_1$ 

0

0

0

1



(b)

Input

 $a_1 \ a_2$ 

0 0

0 1

1 0

1 1

Output

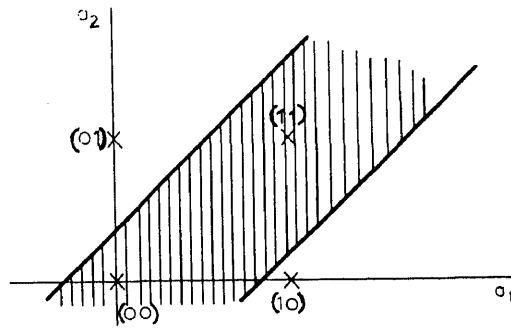
 $b_1$ 

1

0

0

1



**Figure 9.** Two 2-class problems to illustrate linear separability – linearly separable (a) and unseparable (b) cases.

geometrical terms linear separability means that the given set of input patterns  $\{a_i\}$  can be separated into  $M$  distinct regions in the  $N$ -dimensional pattern space by a set of linear hyperplanes. Here  $M$  corresponds to the number of distinct output patterns or classes. As a simple illustration, we can consider 2-dimensional binary (0, 1) patterns in input pattern space and a 1-dimensional output pattern. Two pattern classification problems are shown in figure 9. Note that the number of input patterns (4) is more than the number of dimensions (2) of the input pattern space. These are two-class problems, as the number of distinct outputs are two. Of the two problems in the figure, the first one is linearly separable since a straight line separates the patterns into two regions of desired classes. In the second problem the desired region cannot be obtained by using a single straight line. Note that a straight line is equivalent to a linear hyperplane in a 2-dimensional space.

The restriction of linear separability is due to the function relating input and output patterns. Any arbitrary assignment of an output pattern to a set of input patterns need not result in a linearly separable function, and hence cannot be represented by the two layer network with nonlinear units in the output layer. Thus, although the restriction on the number and type of input patterns (as in the case of pattern association problem) is removed due to introduction of nonlinear units, a restriction is now placed on the nature of the function relating the input and output patterns. To remove this restriction a multilayer feedforward network with nonlinear processing units can be used (Minsky & Papert 1988). Such a network can handle a more general class of pattern classification problems, namely, *pattern mapping* problems which will

be discussed in the next section. Geometrically, it can be argued that a multilayer feedforward neural network can perform classification of patterns with complex boundary surfaces separating different classes in an  $N$ -dimensional space (Lippmann 1987; Minsky & Papert 1988). However, training such a network is not straightforward. Thus it leaves us with a *hard learning problem* which can be solved using the *generalized delta rule* (Rumelhart & McClelland 1986).

**4.1c Pattern mapping:** For a pattern mapping problem the input and output patterns are points in the  $N$ - and  $M$ -dimensional continuous spaces, respectively. The objective is to capture the implied functional relationship or mapping function between the input and output by training a feedforward neural network. This is also called the *generalization problem* (Deuker *et al* 1987). Once the network generalizes by capturing the mapping function through its weights, then during recall from an input pattern the network produces an output which is an interpolated version of the outputs of the training input patterns near the current input pattern. The input patterns are generally naturally occurring patterns as in speech and hand-printed characters.

A multilayer feedforward network with at least two intermediate layers in addition to the input and output layers can perform a pattern mapping task (Cybenko 1989). The number of units in the input and output layers correspond to the dimensions of the input and output patterns, respectively. The additional layers are called *hidden layers*, and the number of units in a hidden layer is determined depending on the problem, usually by trial and error. The network can be trained (i.e. weights at different layers can be adjusted) for a given set of input-output pattern pairs using a *generalized delta rule* or *backpropagation law* (see figure 10) (Rumelhart & McClelland 1986; Hush & Horne 1993). It is derived using the principle of gradient descent along the error surface in the weight space. The given patterns are applied in some random order one by one, and the weights are adjusted using the backpropagation law. The pattern pairs may have to be applied several times till the output error is reduced to an acceptable value.

Once the network is trained, it can be used to recall the appropriate pattern (in this case some interpolated output pattern) for a new input pattern. The computation is straightforward in the sense that the weights and the output functions of the units at different layers are used to compute the activation values and output signals. The signals from the output layer correspond to the output.

Note that for the backpropagation law to work (see figure 10), the output function of the units in the hidden and output layers must be nonlinear and differentiable. Such functions are called *semilinear*. If they are linear, no advantage is obtained by using additional hidden layers. By using a hardlimiting threshold function, it is not possible to propagate the error to hidden layer units to adjust the weights in that layer. Thus the advantage of complex pattern mapping or pattern classification is obtained by a multilayer feedforward network mainly because of the use of the semilinear output functions.

The use of semilinear functions results in a rough error surface in the weight space. That is, there will be several local minima, besides a global minimum. The effects of local minima can be partially reduced by using a stochastic update of weight values (Wasserman 1988). In general the backpropagation learning law needs several iterations in order to reach an acceptably low value of error, at which the network can be assumed to have captured the implied mapping in the given set of input-output

**Backpropagation algorithm: Generalized delta rule**

Given a set of input-output patterns  $\mathbf{a}_l, \mathbf{b}_l, l = 1, 2, \dots, L$

$l$ th input vector  $\mathbf{a}_l = (a_{l1}, a_{l2}, \dots, a_{lN})^T$  and output vector  $\mathbf{b}_l = (b_{l1}, b_{l2}, \dots, b_{lN})^T$

Assume only *one* hidden layer and initial setting of weights to be arbitrary

Assume input layer with only linear units. Then output signal = input activation value

$\eta$  is the learning rate parameter

Activation of unit  $i$  in the input layer  $x_{li} = a_{li}$

Activation of unit  $j$  in the hidden layer  $x_{lj}^h = \sum_{i=1}^N w_{ji}^h x_{li} + \theta_j^h$

Output signal from the  $j$ th unit in the hidden layer,  $s_{lj}^h = f_j^h(x_{lj}^h)$

Activation of unit  $k$  in the output layer  $x_{lk}^o = \sum_{j=1}^H w_{kj}^o s_{lj}^h + \theta_k^o$

Output signal from unit  $k$  in the output layer  $s_{lk}^o = f_k^o(x_{lk}^o)$

Error term for the  $k$ th output unit  $\delta_{lk}^o = (b_{lk} - s_{lk}^o) f_k^{o'}(x_{lk}^o)$

Update the weights on the output layer  $w_{kj}^o(t+1) = w_{kj}^o(t) + \eta \delta_{lk}^o s_{lj}^h$

Error term for the  $j$ th hidden unit  $\delta_{lj}^h = f_j^{h'}(x_{lj}^h) \sum_{k=1}^M \delta_{lk}^o w_{kj}^o$

Update the weights on the hidden layer  $w_{ji}^h(t+1) = w_{ji}^h(t) + \eta \delta_{lj}^h a_{li}$

Calculate the error for the  $l$ th pattern  $E_l = \frac{1}{2} \sum_{k=1}^M \delta_{lk}^2$

Total error for all patterns  $E = \sum_{l=1}^L E_l$

Apply the given patterns, may be several times, in some random order and update the weights until the total error reduces to an acceptable value.

**Figure 10.** Generalized delta rule.

pattern pairs. However, due to the slow rate of convergence of the backpropagation learning law, new architectures (like counter propagation, Hecht-Nielsen 1990) are being sought for faster learning.

## 4.2 Pattern recognition tasks by feedback ANN

**4.2a Autoassociation:** In this section we consider pattern recognition tasks that can be performed by simple feedback neural networks (figure 11). We begin with the autoassociation task discussed earlier when the input and output patterns in each pair are the same i.e.,  $\mathbf{a}_l = \mathbf{b}_l, l = 1, 2, \dots, L$ . The objective in an autoassociation task is to design a network that can recall a stored pattern given a corrupted (noisy or partial) version of the pattern. A feedback network with  $N$  linear processing units can perform the task of autoassociation. Such a network can be trained (i.e., the weights can be determined) using either Hebb's law or delta rule (Hecht-Nielsen 1990). Hebb's learning law leads to a set of optimal weights when the given patterns are orthogonal. Delta rule leads to set of optimal weights when the given patterns are linearly independent.

Pattern recall will be exact when the test pattern is same as one of the stored ones, represented by the weights. If the test pattern is a noisy version of the stored pattern, the recalled pattern is also a noisy version of the stored pattern. In fact the network recalls the input pattern itself, as every vector is associated with itself, thus completely eliminating any accretive behaviour (Murakami & Aibara 1987).

**Auto association (Pattern storage)**

- \* *Arch*: Single layer with feedback, linear processing units
- \* *Learning*: Hebb (orthogonal inputs), Delta (linearly independent inputs)
- \* *Recall*: Direct
- \* *Limitation*: Linear independence of patterns, # of patterns limited to dimensionality
- \* *To Overcome*: Nonlinear processing units, becomes a pattern storage problem

**Pattern storage**

- \* *Arch*: FBNN, nonlinear processing units, states, Hopfield energy analysis
- \* *Learning*: Not important
- \* *Recall*: Activation dynamics until stable states are reached
- \* *Limitation*: False minima, hard problems, limited # patterns
- \* *To Overcome*: Stochastic update, hidden units.

**Pattern environment storage**

- \* *Arch*: Boltzmann machine, nonlinear processing units, hidden units, stochastic update
- \* *Learning*: BM learning law, simulated annealing
- \* *Recall*: Activation dynamics, simulated annealing
- \* *Limitation*: Slow learning
- \* *To Overcome*: Different architecture

**Figure 11.** Pattern recognition tasks by feedback ANN (FBNN).

Thus autoassociation by a feedback network with linear units is not going to serve any purpose. Moreover, the number of patterns is limited to the dimensionality of the pattern. Although there is no simple learning law, it can be shown that the weights of such a network can be determined to store any  $L \leq N$  patterns, without any error in recall, where  $N$  is the dimension of the input pattern space. Discussion of autoassociation task by a feedback network with linear units is only of academic interest, as any input pattern comes out as itself if it is one of the stored ones, and a noise input comes out as a noisy pattern, not as the nearest stored pattern.

To overcome this limitation due to the absence of accretive behaviour, the linear units are replaced with units having nonlinear output functions. The resulting feedback network can then perform *pattern storage* task which will be considered next.

**4.2b Pattern storage:** The objective is to store a given set of patterns so that any one of the patterns can be recalled exactly when an approximate (corrupted) version of the pattern is presented to the network. What is needed is the storage of features and their spatial relations in the patterns, and the pattern recall should take place even when the features and their spatial relations are slightly modified due to noise and distortion. The approximation of pattern refers to the closeness of the features and their spatial relations in the pattern when compared to the original stored pattern. What is actually stored in practice is the information in the pattern data itself. The approximation is measured in terms of some distance, like Hamming distance (in case of binary patterns). The distance feature is automatically realized through the threshold (binary) feature of the output function of a processing unit. The pattern storage is accomplished by a feedback network consisting of nonlinear processing units (see figure 12).

For the simplest case, the weights on the connecting links between units are assumed to be symmetric, i.e.,  $w_{ij} = w_{ji}$ , and that there is no self feedback, i.e.,  $w_{ii} = 0$ . The output signals of all units at any instant of time define the state of the network at that instant. Each state of the network can be assumed to correspond to some *energy*

**Hopfield net algorithm** – To store and recall a set of bipolar patterns

Let the network consist of  $N$  fully connected units with each unit having hard limiting bipolar threshold output function.

Let  $\{a_l\}$ ,  $l = 1, 2 \dots L$  be the vectors to be stored.

The vectors  $\{a_l\}$  are assumed to have bipolar components, i.e.,  $a_{li} = \pm 1$ .

1. Assign the connection weights

$$w_{ij} = \sum_{l=1}^L a_{li} a_{lj}, \quad \text{for } i \neq j$$

$$= 0, \quad \text{for } i = j, \quad 1 \leq i, j \leq L.$$

2. Initialize the network output with the given unknown input pattern  $\mathbf{a}$

$$s_i(0) = a_i, \quad i = 1, 2 \dots N$$

where  $s_i(0)$  is the output of the unit  $i$  at time  $t = 0$ .

3. Iterate until convergence

$$s_i(t+1) = \text{sign} \left[ \sum_{j=1}^N w_{ij} s_j(t) \right], \quad i = 1, 2 \dots N$$

The process is repeated until the outputs remain unchanged with further iteration. The steady outputs of the units represent the stored pattern that best matches the given input.

**Figure 12.** Hopfield Net algorithm to store and recall a set of bipolar patterns.

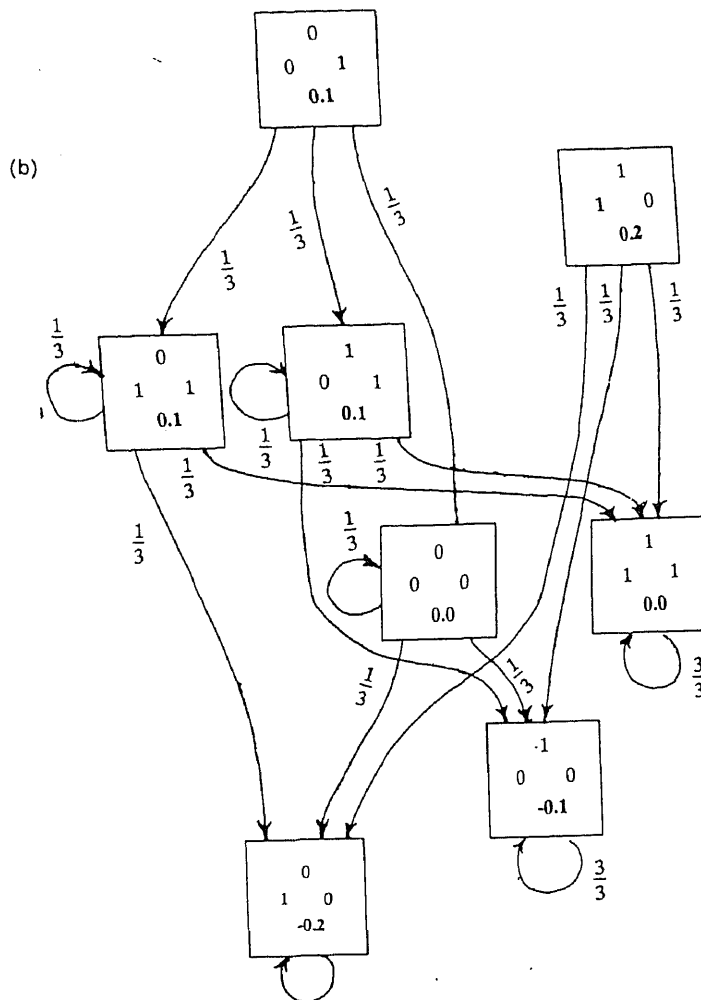
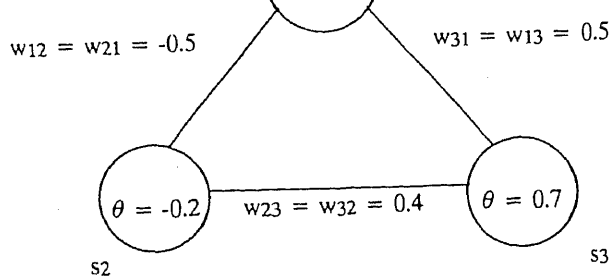
which is defined in terms of the output state  $\{s_i\}$  and weights  $[w_{ij}]$  of the network as (Hopfield 1982)

$$E = -\frac{1}{2} \sum_i \sum_{j \neq i} s_i w_{ij} s_j - \sum_i I_i s_i + \sum_i \theta_i s_i,$$

where  $I_i$  is an external input and  $\theta_i$  is the threshold of the unit. The energy as a function of the output state can be viewed as something like an *energy landscape*. The shape of the landscape is dictated by the network units and their interconnection strengths (weights). The feedback and the nonlinear processing units of the network create *basins of attraction* in the energy landscape. The basins tend to be regions of equilibrium states. If there is a fixed state (point in the output state space) in each of these basins where the energy is minimum, these states corresponds to fixed points of equilibrium. There could also be periodic (or oscillating) regions or chaotic regions of equilibrium (Kosko 1992).

It is the existence of the basins of attraction that is exploited to store the desired patterns and recall them even with approximate inputs as keys. Each pattern is stored at a fixed point of equilibrium of the *energy minimum*. An erroneous or distorted pattern is more likely to be closer to the corresponding true pattern than to the other stored patterns. Each input pattern results in a state of the network that may be closer to the desired state, in the sense that it may lie near the basin of attraction corresponding to the true state. Since an arbitrary state need not correspond to a stable state, the activation dynamics of the network may eventually lead to a *stable state* from which the desired pattern may be read or derived.

If the nonlinear output function of each unit is a binary threshold function (curve A



**Figure 13.** (a) A 3-unit feedback network with symmetric weights and binary threshold units. Activation dynamics  $x_j = \sum_i w_{ji} s_i - \theta_j$ ,  $s_j = f(x_j)$ ; Energy  $E = -\frac{1}{2} \sum_i \sum_j w_{ij} s_i s_j + \sum_i s_i \theta_i$ . (b) State transition diagram for the 3-unit network of figure 13a. Each block represents a state given by sequence  $s_1, s_2, s_3$ . There are eight blocks for eight states. The energy for each state is indicated by the bold numbers with each block. Note that the state diagram has three stable states (111, 100 and 010) (Aleksander & Morton 1990).

in figure 3), then the stable states of the network would lie at the corners of the binary hypercube in the  $N$ -dimensional discrete binary space. On the other hand, if the output function is a semilinear function (curve C in figure 3), then the points corresponding to these states may move closer to each other within the unit hypercube. If the output function is a horizontal line, then almost all states remain close to each other, and hence there will be only one state for the network.

Given a network, it is possible to determine the state transition diagram (Aleksander & Morton 1990). Figure 13 shows the state transition diagram for a 3-unit network. The diagram illustrates the different states of the network and their transition probabilities. States which have self transition with probability 1 are stable states. For a given number of units, the state transition probabilities and the number or stable states are dictated by the connection strengths or weights.

Since each state is associated with some energy value, the state transition diagram shows transitions from a state with higher energy value to a state having lower or equal energy value. The energy value of a stable state corresponds to an energy minimum in the landscape, as there is no transition from this to the other states.

The number of basins of attraction in the energy landscape depends only on the network, i.e., the number of processing units and their interconnection strengths (weights). When the number of patterns to be stored is less than the number of basins of attraction, i.e., stable states, then there will be spurious stable states, which do not correspond to any desired patterns. In such a case, when the network is presented with an approximate pattern for recall, the activation dynamics may eventually lead to a stable state which may correspond to one of the spurious states or a *false energy minimum*, or to one of the stable states corresponding to some other pattern. In the latter case there will be an undetected error in the recall. The average probability of error depends on the energy values of the stable states corresponding to the desired patterns, and the relative locations of these states in the state space, measured in terms of some distance criterion.

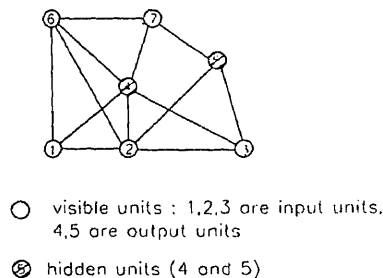
If the number of desired patterns to be stored is more than the number of basins of attraction in the energy landscape, then the problem becomes a *hard problem*, in the sense that the given patterns cannot be stored in the network.

For a given network it is not normally possible to determine exactly the number of basins of attraction as well as their relative spacings and depths in the state space of the network. It is possible to estimate the *capacity* (number of patterns that can be stored) of the network and also the average *probability of error* in recall (Abu-Mostafa & St. Jaques 1985; Aleksander & Morton 1990). The probability of error in recall can be reduced by adjusting the weights in such a way that the resulting energy landscape is matched to the probability distribution of the input patterns. This becomes the problem of storing a *pattern environment*.

**4.2c Pattern environment storage:** A pattern environment is described by the set of desired patterns together with their probability distribution. The objective is to store a pattern environment in a network in such a way that the average probability of error in recall is minimized. This is achieved if the energy landscape is designed in such a way that the desired patterns are stored at the stable states corresponding to the lowest minima, with the higher probability patterns at lower energy minima points.

Boltzmann machine architecture together with the Boltzmann learning law can achieve an optimal storage of pattern environment (Hinton & Sejnowski 1986;





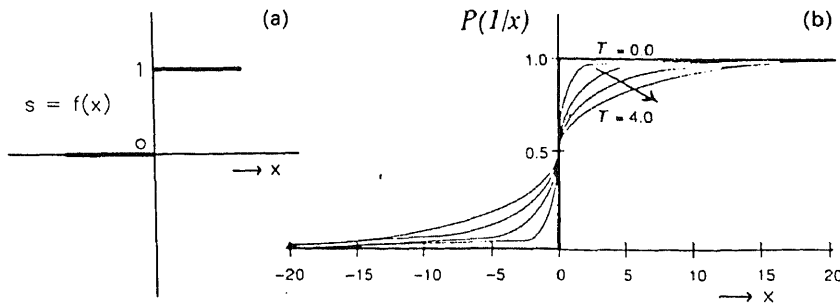
**Figure 14.** Architecture of the Boltzmann machine. Each unit is connected to every other unit, although only a few connections are shown in the figure. Some of the visible units can be identified as input units and others as output units if the machine is to be used for pattern mapping.

Aleksander & Morton 1990). The architecture consists of a number of processing units with each unit connecting to all the other units (figure 14). The number of units is typically larger than the dimension of the input pattern. The additional units are called *hidden units*. Use of hidden units helps in overcoming the limitation of the hard problems of pattern storage by a fully connected network. The patterns are applied to the so-called *visible units*, the number of visible units being equal to the dimension of the input patterns.

Error in pattern recall due to false minima can be reduced significantly if initially the desired patterns are stored (by careful training) at the lowest energy minima. The remaining error can be reduced by using suitable activation dynamics. Let us assume that by training we have achieved a set of weights which will enable the desired patterns to be stored at the lowest energy minima. The activation dynamics is modified so that the network can also move to a state of higher energy value initially, and then to the nearest deep energy minimum. It is possible to realize this by using a *stochastic update* in each unit instead of the deterministic update of the output function as in the previous cases. By stochastic update we mean that the activation value or the net input to a unit need not decide the next output state of the unit in a deterministic manner as in the case of figure 12. The update is expressed in probabilistic terms, like the probability of firing the unit being greater than 0.5 if the net input exceeds a threshold, and less than 0.5 if the net input is less than the threshold for the unit. Note that the output function could still be a threshold logic (hardlimiter), but it is applied in a stochastic manner.

With the new activation dynamics, the state transition diagram shows transitions from a lower energy state to a higher energy state as well, the probability of such a transition is dictated by the probability function used in determining the firing of a unit in the stochastic update (Aleksander & Morton 1990). The probability function (figure 15) can in turn be defined in terms of a parameter, called *temperature* ( $T$ ). As the temperature is increased, the uncertainty in the update increases, giving the network a greater chance to go to a higher energy level state.

Since eventually we want the activation dynamics to lead the network to a stable state corresponding to the pattern closest to the given input pattern, we need to provide greater mobility for transition to higher states only initially. The mobility is slowly decreased by reducing the temperature, eventually to  $T=0$ . At the lowest temperature the network settles down to a fixed point state corresponding to the desired pattern. At each temperature the network dynamics is allowed to settle to some equilibrium situation, called *thermal equilibrium*. At thermal equilibrium the average probability of visiting the states of the networks will not change further. The temperature parameter is reduced in a predetermined manner (called annealing



**Figure 15.** Stochastic update of a unit using probability law. Probability of firing  $P(1/x) = 1/[1 + \exp(-x/T)]$ . (a) Binary output function; (b) probability function for stochastic update.

schedule), making sure that at each temperature the network is allowed to reach thermal equilibrium before the next change in temperature is made. This process is called *simulated annealing* (Kirkpatrick *et al* 1983) (see figure 16). Note that at each temperature the state update dynamics is fixed, as the probability of transition from one state to another depends only on the temperature. The update dynamics is however altered when the temperature is changed, and it results in a new state transition diagram.

The states at thermal equilibrium at  $T = 0$  represent the stable states of the network corresponding to the minima of the energy landscape. The probabilities of these states are also related to the actual minimum energy values of the states. The relation between the probabilities of stable states and energy suggest that the probability of error in the recall of patterns can be further reduced if the probability distribution of the desired patterns, i.e., the pattern environment, is known, and is used in determining the optimal setting of weights of the network.

---

#### Simulated annealing algorithm – To recall a stored pattern with partial input

Let us assume a Boltzmann machine with some visible units and some hidden units.

Let the network consist of total  $N$  fully connected units, with each unit having a hard limiting binary threshold output function. Let us assume that the network was already trained to store the given set of input patterns.

1. Force the outputs of the visible units to the corresponding known components in the given partial binary input vector.

2. Assign for all unknown visible units and all hidden units to random binary output values.

3. Select a unit  $k$  at random, and calculate its activation value  $x_k$  using weighted sum of its inputs.

4. Assign the output of the unit  $k$  to 1 with probability  $P_k = \frac{1}{1 + \exp(-x_k/T)}$ , where  $T$  is the temperature parameter.

5. Repeat steps 3 and 4 until all units have had the same probability of being selected for update. This number of unit-updates defines a processing cycle.

6. Repeat step 5 for several processing cycles until *thermal equilibrium* has been reached at the given temperature  $T$ , i.e., when the probability of visiting different states of the network does not change any further. This is usually accomplished only approximately.

7. Lower the temperature, and repeat steps 3 through 7 until a stable state is reached at which point there will not be any further change in the state of the network. The result of recall is the stable output state of the visible units.

---

**Figure 16.** Simulated annealing algorithm.

### Learning in Boltzmann machine

The objective is to adjust the weights of a Boltzmann machine so as to store a pattern environment described by the set of vectors  $\{V_a\}$  and their probabilities of occurrence. These vectors should appear as the outputs of the visible units. Define  $\{H_b\}$  as the set of vectors appearing on the hidden units.

Let  $P^+(V_a)$  be the probability that the outputs of the visible units will be clamped (indicated by “+” superscript) to the vector  $V_a$ . Then,

$$P^+(V_a) = \sum_b P^+(V_a \wedge H_b),$$

where  $P^+(V_a \wedge H_b)$  is the probability of the state of the network when the outputs of the visible units are clamped to the vector  $V_a$ , and the outputs of the hidden units are  $H_b$ .

Likewise the probability that  $V_a$  will appear on the visible units when none of the visible units are clamped (indicated by “-” superscript) is given by

$$P^-(V_a) = \sum_b P^-(V_a \wedge H_b).$$

Note that  $P^+(V_a)$  is given by the pattern environment description, and  $P^-(V_a)$  depends on the network dynamics and is given by

$$P^-(V_a) = \sum_b \exp(-E_{ab}/T) / \sum_{m,n} \exp(-E_{mn}/T),$$

where the total energy of the system in the state  $V_a \wedge H_b$  is given by

$$E_{ab} = -\frac{1}{2} \sum_{i,j} w_{ij} s_i^{ab} s_j^{ab},$$

$s_i^{ab}$  refers to the output of the  $i$ th unit in the state  $V_a \wedge H_b$ .

The Boltzmann learning law is derived using the negative gradient descent of the functional

$$G = \sum_a P^+(V_a) \ln[P^+(V_a)/P^-(V_a)].$$

It can be shown that

$$-\partial G / \partial w_{ij} = (1/T)(P_{ij}^+ - P_{ij}^-),$$

where

$$P_{ij}^+ = \sum_{a,b} P^+(V_a \wedge H_b) s_i^{ab} s_j^{ab},$$

$$P_{ij}^- = \sum_{a,b} P^-(V_a \wedge H_b) s_i^{ab} s_j^{ab}.$$

The weight updates are calculated according to

$$\Delta w_{ij} = -\eta(\partial G / \partial w_{ij}) = \eta(1/T)(P_{ij}^+ - P_{ij}^-).$$

The Boltzmann law is implemented using some annealing schedule for the network during clamped and unclamped phases of the visible units of the network to determine  $P_{ij}^+$  and  $P_{ij}^-$ , respectively.

Figure 17. Boltzmann learning law.

The Boltzmann learning law (see figure 17) allows us to represent a given environment by the network (Ackley *et al* 1985; Hinton & Sejnowski 1986; Aleksander & Morton 1990). The law uses an information theoretic measure to evaluate how well the environment is represented in the network. If a perfect representation is obtained, then there will be as many energy minima as there are desired patterns. But in practice only an approximate representation of the environment is accomplished, and hence there will be some spurious stable states which correspond to the false wells in the energy landscape. The Boltzmann learning law uses a simulated annealing schedule for implementation, i.e., for determining the weight updates at each stage. Recall of stored patterns from an approximate input pattern also uses a simulated annealing schedule to overcome the false minima created because of the approximate representation of the environment by the network.

In general the Boltzmann learning law converges slowly to the desired weights (Geman & Geman 1984; Szu 1986, pp. 420–5). Moreover, there is no simple way to determine the optimum number of hidden units for a network to solve the given problem of pattern environment storage. The larger the number of hidden units, the greater is the chance for more false minima, and hence the greater the probability of error in recalling a stored pattern. The smaller the number of hidden units, the greater the chance that the given problem becomes hard for the network. New architectures are needed to overcome some of these limitations of the Boltzmann machine for the problem of pattern environment storage.

#### 4.3 Pattern recognition tasks by feedforward and feedback ANN

In this section we discuss some pattern recognition tasks (figure 18) that can be performed by a network consisting of two layers of processing units: The first layer with linear output units feeds the input pattern to the units in the second layer through a set of feedforward connections with appropriate weights. The outputs of

---

##### Pattern storage (STM)

- \* *Arch*: Two layers (input & competitive), linear processing units
- \* *Learning*: No learning in FF stage, fixed weights in FB layer
- \* *Recall*: Not relevant
- \* *Limitation*: STM, no application, theoretical interest
- \* *To overcome*: Nonlinear output function, learning in FF stage

##### Pattern clustering (grouping)

- \* *Arch*: Two layers (input & competitive), nonlinear processing units
- \* *Learning*: Only in FF stage – Competitive learning
- \* *Recall*: Direct, activation dynamics until stable state is reached
- \* *Limitation*: Fixed (rigid) grouping of patterns
- \* *To overcome*: Neighbourhood units in competition layer

##### Feature map

- \* *Arch*: Self-organization network, 2 layers, nonlinear processing units
  - \* *Learning*: Neighbourhood units in competitive layer
  - \* *Recall*: Apply input, determine winner
  - \* *Limitation*: Only visual features, not quantitative
  - \* *To overcome*: More complex architecture
- 

**Figure 18.** Pattern recognition tasks by feedforward (FF) and feedback (FB) ANN.

the units in the second layer are fed back to the units in the same layer including feedback to the same unit. The self feedback is usually with a positive weight (excitatory connection) and the feedback to the other units is usually with a negative weight (inhibitory connection). The weights on the feedback connections in the second layer are usually fixed. The first layer of units is called input layer, and the second layer is called *competitive layer* (Rumelhart & Zipser 1986). Different choices of output functions and methods of learning lead to networks for different types of competition tasks. We discuss three such tasks. Assuming fixed weights in the feedforward connections from the input to the competitive layer, and in the feedback connections in the competitive layer, we can study the behaviour of the network for different types of output functions of the units in the competition layer.

**4.3a Pattern storage (short term memory):** First let us assume the output functions to be linear. When an input pattern is applied, the units in the competition layer settle to a steady activation state which will remain there even after the input pattern is removed (Freeman & Skupura 1991). The activation pattern will remain as long as the network is not given a different input pattern. Another input pattern will erase the previous activation state. Hence this is called short-term memory. The pattern is stored only temporarily.

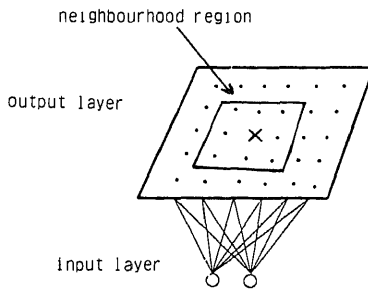
This pattern storage representation is only of theoretical interest. There is no application for such a short-term memory function. However, by using a nonlinear output function for the units in the competition layer one could show that the network can perform a *pattern clustering* task.

**4.3b Pattern clustering:** Given a set of patterns, the objective is to design a competition network which groups the patterns into subgroups of patterns based on similarity of features in the patterns. A two layer network with input and competition layers, and with nonlinear units in the competition layer can perform the task of pattern clustering or grouping (Grossberg 1980).

If a nonlinear output function of the type  $f(x) = x^2$  is used for the units in the competitive layer, then it can be shown (Freeman & Skupura 1991) that the activation dynamics leads to a steady state situation where the network tends to enhance the activity of the unit with the largest activity. When the input pattern is removed, the activities of all units except the largest one will decay to zero. Thus only one of the units in the competitive layer will win. The weights leading to the winning unit  $j$  are adjusted to respond more to the input pattern  $\mathbf{a}$ . This weight adjustment is repeated for all the input patterns several times. For input patterns belonging to different groups, different units in the competition layer will win. When the weight vector for each output unit reaches an average position within the cluster, it will stay generally within a small region around that average position. Each unit in the competitive layer refers to a different group or category of patterns.

When an unknown input pattern is given, the activation dynamics leads to a steady state situation where only one unit in the competitive layer is active. That unit gives the category to which the input pattern belongs.

Note that in a competitive network the physical location of the units do not reflect any relation between categories. But there are many situations where the patterns do not fall into fixed categories. There may be a gradual change of features from one pattern to another. This change of features can be captured by a selforganisation network which performs the task of *feature mapping* (von der Malsburg 1973; Willshaw & von der Malsburg 1976).



**Figure 19.** A feature mapping architecture. Each input unit is connected to all the units of the output layer.

**4.4b Feature map:** Given a set of patterns, the objective is to design a network that would organise the patterns in accordance with similarity of features among them in such a way that by looking at the output of the network one can visually obtain an idea of how different patterns are related. The display of signals from the output layer (typically in 2-dimension) of units is called a *feature map*.

To accomplish the task of feature mapping, a competitive network is modified into one called a *selforganising network* (Kohonen 1990; Freeman & Skupura 1991) shown in figure 19. The modification consists of creating a neighbourhood region around the winning unit in the competitive layer, so that during training all the feedforward weights leading to the units in this region are adjusted to favour the input pattern. The weight adjustment is similar to the case of a competitive network. The neighbourhood region around a winning unit is gradually reduced for each application of the given set of patterns (see figure 20).

---

#### Algorithm for self-organizing feature map

1. Initialize the weights from  $N$  inputs to the  $M$  output units to small random values. Initialize the size of the neighbourhood region  $R(0)$ .
2. Present a new input  $\mathbf{a}$
3. Compute the distance  $d_i$  between input and the weight on each output unit  $i$ :

$$d_i = \sum_{j=1}^N [a_j(t) - w_{ij}(t)]^2, \text{ for } i = 1, 2, \dots, M$$

where  $a_j(t)$  is the input to the  $j$ th input unit at time  $t$  and  $w_{ij}(t)$  is the weight from the  $j$ th input unit to the  $i$ th output unit.

4. Select the output unit  $i^*$  with minimum distance

$$i^* = \text{index of } \left[ \min_i (d_i) \right]$$

5. Update weight to node  $i^*$  and its neighbours

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(a_j(t) - w_{ij}(t)), \\ \text{for } i \in R_i^*(t), \text{ and } i = 1, 2, \dots, N,$$

where  $\eta(t)$  is the learning rate parameter ( $0 < \eta(t) < 1$ ) that decreases with time  $R_i^*(t)$  gives the neighbourhood region around the node  $i^*$ , at time  $t$ .

6. Repeat steps 2 through 5 for all inputs several times.
- 

**Figure 20.** An algorithm for self-organizing feature map.

For recall, when an unknown input is applied, the activation dynamics determines the winning unit whose location would determine its features relative to the features represented by the other units in its neighbourhood.

While a feature map produces a more realistic arrangement of patterns, the output is useful only for visual observation. Since it is difficult to categorize a feature map, it is difficult to use it for applications such as pattern classification. A more complex architecture is needed to exploit the advantages of a feature map for pattern classification purposes (Huang & Kuh 1992).

## 5. Architectures for complex pattern recognition tasks

So far we have considered simple structures of neural networks and discussed the pattern recognition tasks that these structures could accomplish. In practice the pattern recognition tasks are much more complex, and each task may require evolving a new structure based on the principles discussed in the previous sections. In fact designing an architecture for a given task involves developing a suitable structure of the neural network and defining appropriate activation and synaptic dynamics. In this section we will discuss some general architectures for complex pattern recognition tasks.

### 5.1 *Associative memory: pattern storage – BAM*

Pattern storage is the most obvious pattern recognition task that one would like to accomplish by an ANN. This is a memory function, where the network is expected to store the pattern information for later recall. The patterns to be stored may be spatial or spatiotemporal (pattern sequence). Typically an ANN behaves like an associative memory, in which a pattern is associated with another, or with itself. This is in contrast with the random access memory which maps an address to a data. An ANN can also function as a content addressable memory where data are mapped to an address.

The pattern information is stored in the weight matrix of a feedback neural network. The stable states of the network represent the stored patterns, which can be recalled by providing an external stimulus in the form of partial input. If the weight matrix stores the given patterns, then network becomes an autoassociative memory. Several architectures are proposed in the literature for realizing an associative memory function depending on whether the pattern data is discrete/continuous, or the network is operating in discrete time/continuous time, or the learning is taking place off-line/on-line (Simpson 1990).

We will discuss the discrete bidirectional associative memory (BAM) in some detail. It is a two-layer heteroassociative neural network (figure 21) that encodes arbitrary binary spatial patterns using Hebbian learning. It learns on-line and operates in discrete time. The BAM weight matrix is given by,

$$W = \sum_{l=1}^L \mathbf{a}_l^T \mathbf{b}_l$$

where  $\mathbf{a}_l \in \{-1, +1\}^N$  and  $\mathbf{b}_l \in \{-1, +1\}^N$ . The superscript  $T$  refers to the transpose of the vector.

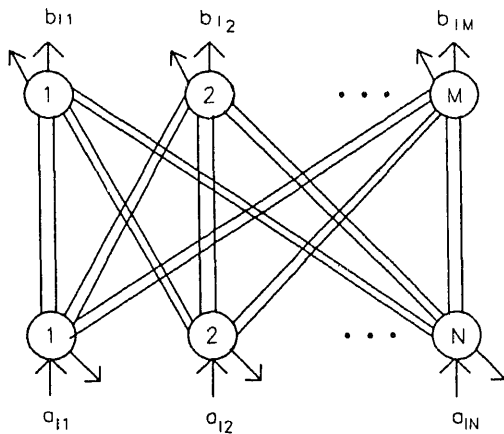


Figure 21. Discrete bidirectional associative memory.

The activation equations are as follows:

$$b_j(t+1) = \begin{cases} 1, & \text{if } y_j > 0, \\ b_j(t), & \text{if } y_j = 0, \\ -1, & \text{if } y_j < 0, \end{cases} \quad \text{where } y_j = \sum_{i=1}^N a_i(t)w_{ji},$$

$$a_i(t+1) = \begin{cases} 1, & \text{if } x_i > 0, \\ a_i(t), & \text{if } x_i = 0, \\ -1, & \text{if } x_i < 0, \end{cases} \quad \text{where } x_i = \sum_{j=1}^M b_j(t)w_{ij}.$$

For recall, the given input  $a_i(0)$ ,  $i = 1, 2, \dots, N$  is applied and the stable values of  $b_j(\infty)$ ,  $j = 1, 2, \dots, M$  are read out. BAM updates are synchronous in the sense that the units in each layer are updated simultaneously.

BAM can be shown to be unconditionally stable (Kosko 1988). However its storage is limited to a small number of binary/bipolar patterns.

## 5.2 Pattern mapping: Data compression – CPN

In pattern mapping the objective is to capture the implied functional relationship between an input–output vector pair  $(\mathbf{a}_i, \mathbf{b}_i)$ . We have seen earlier that a multilayer feedforward network with a semilinear output function can perform generalization, but the training process is slow, and the ability to generalize depends on the learning rate and the number of units in the hidden layers. Several architectures are proposed in literature for realizing a mapping function (Simpson 1990). A practical approach for implementing pattern mapping is to use an architecture that learns fast. A counter-propagation network (CPN) that uses a combination of instar and outstar topologies is proposed (figure 22) for this purpose (Hecht-Nielson 1987). It consists of a three-layer feedforward network with the first two layers forming a competitive learning system and the second (hidden) and third layers forming an outstar structure. Learning takes place in the instar structure of the competitive learning system to code the input patterns  $\{\mathbf{a}_i\}$  and in the outstar structure to represent the output patterns  $\{\mathbf{b}_i\}$ . The training of the instar and outstar structures are as follows.



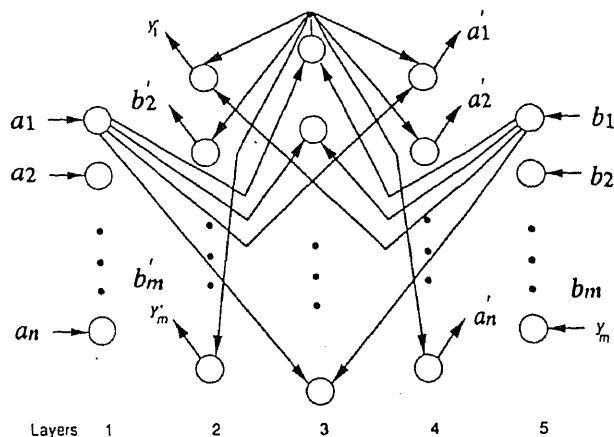


Figure 22. Counter propagation network.

### Training instars of CPN

- (1) Select an input vector  $\mathbf{a}_l$  from the given training set  $(\mathbf{a}_l, \mathbf{b}_l)$ ,  $l = 1, 2, \dots, N$ .
- (2) Normalize the input vector and apply it to the CPN competitive layer.
- (3) Determine the unit that wins the competition by determining the unit  $m$  whose vector  $\mathbf{w}$  is closest to the given input.
- (4) Update the winning unit's weight vector as

$$\mathbf{w}_m(t+1) = \mathbf{w}_m(t) + \eta(\mathbf{a}_l - \mathbf{w}_m(t)).$$

- (5) Repeat steps 1 through 4 until all input vectors are grouped properly by applying the training set several times.

After successful training the weight vectors leading to each hidden unit represent the average of the input vectors corresponding to the group represented by the unit.

### Training outstars of CPN

- (1) After training instars, apply a normalized input vector  $\mathbf{a}_l$  to the input layer and the corresponding output  $\mathbf{b}_l$  to the output layer.
- (2) Determine the winning unit  $m$  in the competitive layer.
- (3) Update the weights on the connections from the winning competitive unit to the output units

$$\mathbf{w}_m(t+1) = \mathbf{w}_m(t) + \beta(\mathbf{b}_l - \mathbf{w}_m(t)).$$

- (4) Repeat steps 1 through 3 until all the vector pairs in the training set are mapped satisfactory.

After successful training, the outstar weights for each unit in the competitive layer represents the average of the subset of the output vectors  $\mathbf{b}_l$  corresponding to the input vectors belonging to that unit.

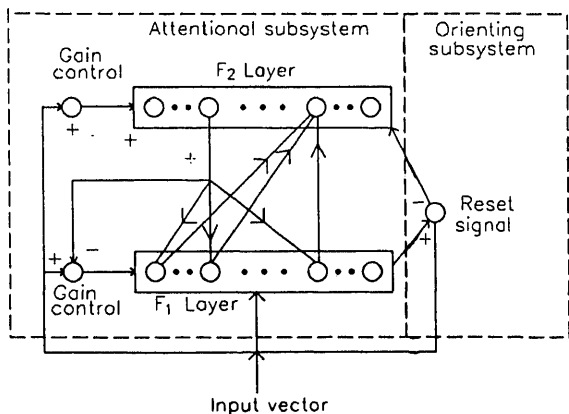
Depending on the number of nodes in the hidden layer, the network can perform any desired mapping function. In the extreme case, if a unit is provided in the hidden layer for each input pattern, then any arbitrary mapping  $(\mathbf{a}_l, \mathbf{b}_l)$  can be realized. But in such a case the network fails to generalize. It merely stores the pattern pair. By using a small number of units in the hidden layer, the network can accomplish data compression. Note also that the network can be trained to capture the inverse mapping

as well, i.e.,  $\mathbf{a}_i = \phi^{-1}(\mathbf{b}_i)$ , provided such a mapping exists and it is unique. The name counterpropagation is given due to the network's ability to learn both forward and inverse mapping functions.

### 5.3 Pattern classification: stability-plasticity dilemma – ART

Many pattern mapping networks can be transformed to perform pattern classification or category learning tasks. However these networks have the disadvantage that during learning the weight vectors tend to encode the presently active pattern, thus weakening the traces of patterns it had already learnt. Moreover any new pattern that does not belong to the categories already learnt, is still forced into one of them, using the best match strategy without taking into account how good even the best match is. The lack of stability of weights as well as lack of inability to accommodate patterns belonging to new categories, led to the proposal of new architectures for pattern classification. These architectures are based on adaptive resonance theory (ART) and are specially designed to take care of the so called *stability-plasticity dilemma* in pattern classification (Carpenter & Grossberg 1988).

ART also uses a combination of instar-outstar network as in CPN, but with the output layer merged with the input layer, thus forming a two-layer network with feedback as shown in figure 23. The minimal ART network includes a bottom-up competitive learning system ( $F_1$  to  $F_2$ ) combined with a top-down ( $F_2$  to  $F_1$ ) outstar pattern learning system. The number of units in the  $F_2$  layer determines the number of possible categories of input patterns. When an input pattern  $\mathbf{a}_i$  is presented to the  $F_1$  layer, the system dynamics initially follows the course of competitive learning, leading to a winning unit in the competitive  $F_2$  layer depending on the past learning of the adaptive weights of the bottom-up connections from  $F_1$  to  $F_2$ . The signals are sent back from the winning unit in the  $F_2$  layer down to  $F_1$  via a top-down outstar network. The activation values produced in the units of  $F_1$  due to this feedback are compared with the activation values due to input. If the two activation patterns match well, then the winning unit in the  $F_2$  layer determines the category of the input pattern. If the match between activations due to top-down and input pattern is poor, as determined by a vigilance parameter, then the winning unit in  $F_2$  does not represent the proper class for the input pattern  $\mathbf{a}_i$ . That unit is removed from the set of allowable winners in the  $F_2$  layer. The other units in the  $F_2$  layer are likewise skipped until a



**Figure 23.** Adaptive resonance theory (ART) architecture. Two major subsystems are the attentional subsystem and the orienting subsystem. Units in each layer are fully interconnected to the units in the other layer.

suitable match is obtained between the activations in the  $F_1$  layer due to top-down pattern and the input pattern. When a match is obtained, then both the bottom-up and top-down network weights are adjusted to reinforce the input pattern. If no match is obtained then an uncommitted (whose category is not identified during training) unit in the  $F_2$  layer is committed to this input pattern, and the corresponding weights are adjusted to reinforce the input.

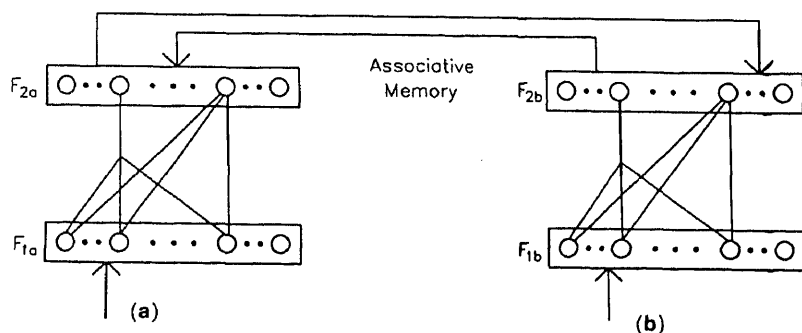
The above sequence of events conducts a search through the encoded patterns associated with each category trying to find a sufficiently close match with the input pattern. If no category exists, a new category is made. The search process is controlled by two subsystems, namely the orienting subsystem and the attentional subsystem. The orienting subsystem uses the dimensionless vigilance parameter that establishes the criterion for deciding whether the match is good enough to accept the input pattern as an exemplar of the chosen category. The gain control process in the attentional subsystem allows the units in  $F_1$  to be engaged only when an input pattern is present, and it also actively regulates the learning (Freeman & Skupura 1991).

Stability is achieved in the ART network due to the dynamic matching and the control in learning. Plasticity is achieved in the ART due to its ability to commit an uncommitted unit in the  $F_2$  layer for an input pattern belonging to a category different from what was already learnt.

ART gets its name from the particular way in which learning and recall interplay in the network. Information in the form of output signals from units reverberate back and forth between the two layers. If the proper patterns develop, a stable oscillation ensues, which is the neural network equivalent of resonance. During this resonance period learning or adjustment of adaptive weights takes place. Before the network has achieved a resonant state, no learning takes place, because the time required for changes in the weights is much longer than the time it takes the network to achieve resonance.

ART1 network was proposed to deal with binary input patterns (Carpenter & Grossberg 1988). ART2 network was developed to selforganize recognition categories for analog as well as binary input patterns (Carpenter & Grossberg 1987).

A minimal ART network can be embedded in a larger system to realize an associative memory. A system like CPN or multilayer perceptron directly maps pairs of patterns ( $a_i, b_i$ ) during learning. If an ART system replaces the CPN, the resulting system becomes self stabilizing. Two ART systems can be used to pair sequences of the categories selforganized by the input sequences as shown in figure 24. The pattern recall can



**Figure 24.** Two ART system combined to form an associative memory architecture.

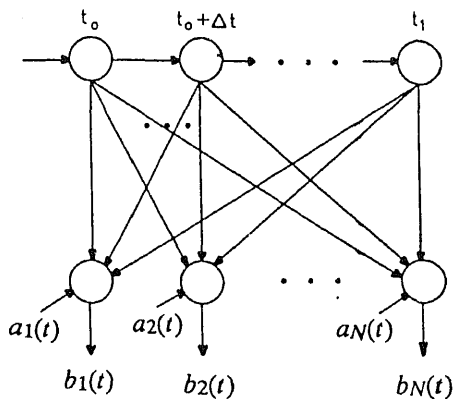


Figure 25. Grossberg's formal avalanche architecture.

occur in either direction during performance as in BAM. This scheme brings to the associate memory paradigm the code compression capabilities, as well as the stability properties of ART (Carpenter 1989).

#### 5.4 Spatio-temporal patterns: temporal features – Avalanche

The ANN architectures described so far are applicable for recognition of patterns on the basis of information contained within the pattern itself. Even if a sequence of patterns with temporal correlation are presented, the previous or subsequent patterns have no effect on the classification of the current input pattern. But there are many applications (for example, speech recognition) where it is necessary to encode the information relating to the time correlation of spatial patterns, as well as the spatial pattern information itself.

Architectures for classification of spatio-temporal patterns (STP) are based on the Grossberg formal avalanche structure (Grossberg 1969). The structure (figure 25) of the network resembles the top two layers of the CPN, and both use multiple outstars. The avalanche architecture shows how a complex spatio-temporal pattern can be learned and recalled. Assume  $\mathbf{a}(t) = (a_1(t), a_2(t), \dots, a_N(t))$  the spatial pattern required at time  $t$ . The sequence of  $\mathbf{a}(t)$  at time intervals of  $\Delta t$  in the range  $t_0 \leq t \leq t_1$  correspond to the desired spatio-temporal pattern. Activate the node labelled  $t_0$  and apply  $\mathbf{a}(t_0)$  to be learned by the outstar's output units. The second pattern  $\mathbf{a}(t + \Delta t)$  is applied while activating the second outstar, labelled  $t_0 + \Delta t$ . Continue this process by activating successive outstars until all the patterns have been learned in sequence. Replay of the learned sequence can be initialized by stimulating the  $t_0$  node, while a zero vector is applied to the  $\mathbf{a}$  inputs. The output sequence  $\mathbf{b}(t) \approx \mathbf{a}(t)$ , for  $t_0 \leq t \leq t_1$ , is the learned sequence.

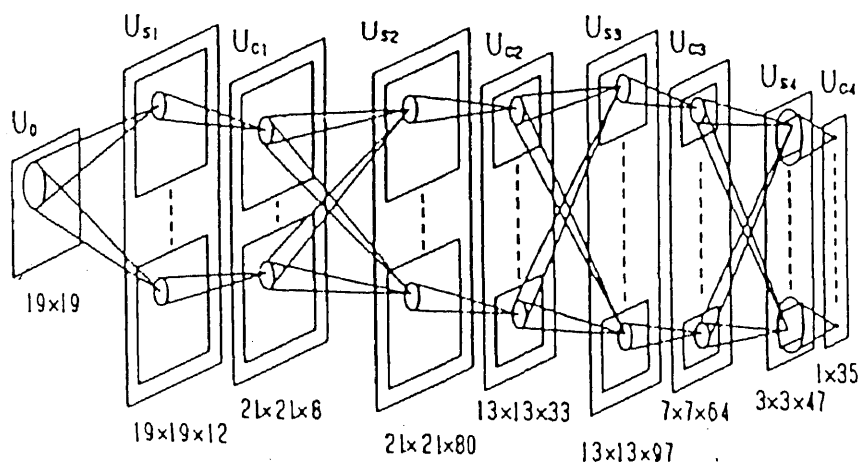
#### 5.5 Pattern variability: recognition of deformed patterns – Neocognitron

Visual pattern recognition, such as recognition of handwritten characters or hand-drawn figures, is done effortlessly by human beings despite variability of features in different realizations of the pattern of the same character or figure. The patterns considered in the architectures described so far assume that the objects in the training and test patterns are identical in size, shape and position, except that in some cases there may be some noise added or some portions of the pattern missing. Models of

associative memory can recover complete patterns from such imperfections, but normally cannot work if there is variability or deformation in the patterns of the test input.

Neural network models based on our understanding of human visual pattern recognition tend to perform well even for shifted and deformed patterns. In the visual area of the cerebrum, neurons respond selectively to local features of a visual pattern such as lines and edges. In areas higher than the visual cortex, cells exist that respond selectively to certain figures like circles, triangles, squares, human faces etc (Fukushima 1975). Thus the human visual system seems to have a hierarchical structure in which simple features are first extracted from the stimulus pattern, then integrated into more complicated ones. A cell at a higher stage generally receives signals from a wider area of the retina and is less sensitive to the position of the stimulus. Within the hierarchical structure of the visual systems are forward (afferent or bottom-up) and backward (efferent or top-down) propagation of signals. This kind of physiological evidence suggests a neural network structure for modelling the phenomenon of visual pattern recognition.

The objective is to synthesize a neural network model for pattern recognition for shifted and deformed patterns. The network model learns with a teacher (supervised learning) for reinforcement of the adaptive weights. The network model is called neocognitron. It is a hierarchical network (figure 26) consisting of many layers of cells, and has variable connections between cells in adjoining layers. It can be trained to recognize any set of patterns. After training, pattern recognition is performed on the basis of similarity in shape between patterns, and the recognition is not affected by deformation, or changes in size, or shifts in the positions of the input patterns (Fukushima 1988).



**Figure 26.** A hierarchical network structure of neocognitron (Fukushima 1991) for recognition of alphanumeric character recognition. The lowest stage of the network consists of a 2-dimensional array of receptor cells. Each succeeding stage has a layers consisting of S cells and C cells alternatively. Each layer is organized into groups of these cells, each group responding to a particular geometrical position. The numbers show the total numbers of S and C cells in individual layers of the network. S cells are feature extracting cells. The C cells are inserted to allow for positional errors in the feature.

In the hierarchical network of the neocognitron, local features of the input pattern are extracted by the cells of the lower stage, and they are gradually integrated into more global features. Finally, each cell of the highest stage integrates all the information of the input pattern, and responds only to one specific pattern. During the process of extracting and integrating features, errors in the relative positions of the local features are gradually tolerated. The operation of tolerating positional error a little at a time at each stage, rather than all in one step, plays an important role in endowing the network with the ability to recognize even distorted patterns (Fukushima 1991).

Neocognitron also provides backward connections which will enable it to realize the selective attention feature of the visual pattern recognition. The selective attention feature relates to two or more patterns simultaneously present in the data, and our ability to focus on the desired one.

Neocognitron was developed for recognition of handwritten characters, although the ideas used in the architecture may be extended to other situations of pattern variability (Fukushima 1991).

## 6. Applications

In this section of the paper we briefly discuss the application potential of neural network models and some research issues that are being currently addressed in this field. In applications we consider two different situations, one where the existing neural network concepts can be directly applied, and the other where there is potential for applying the neural network ideas but it is not yet clear how to formulate the real world problems to evolve a suitable neural network architecture. We will also list a few cases where neural network principles are being used in practice.

### 6.1 Direct application

In applications such as associative memories, optimization, vector quantization and pattern classification the principles of neural networks are directly applicable. In these applications it is assumed that the problem can be presented to the network directly, and what is being sought is the solution to the problem using the dynamics of the network. Many real world problems were formulated into one of these, and were solved successfully (Lisboa 1992).

6.1a *Associative memories* (Bienenstock & von der Malsburg 1987; Hassoun 1989; Desai 1990; Kamp & Hasler 1990; Michel & Farrell 1990): As discussed earlier, the objective of associative memory is to store a pattern or data for later recall with partial or noisy version of the pattern as input, or to store association between two patterns for later recall of one of the patterns given the other. Both feedback and feedforward topologies of neural networks are directly used for these applications. Associative memory, if used in a feedback structure of the Hopfield type, can function as a content addressable memory as well. The stable states of the network, which represent the energy minima or basins of attraction, are used to store the pattern information. In a feedforward network the mapping function corresponding to the input-output pattern pairs is stored in the weights of the network.

Applications of these networks for associative memory is direct, if the patterns are

available in the form of one or two dimensional array of values. Associative memories as content addressable memories are quite powerful. For example, if information about individuals is stored in a network, then it is possible to retrieve the complete data by providing partial or even noisy clues. Other common applications for an associative memory are recognition of images, and retrieval of bibliography information from partial references such as from incomplete title of a paper.

**6.1b Optimization:** One of the most successful applications of neural network principles is in solving optimization problems (Hopfield & Tank 1985; Kennedy & Chua 1988; Rauch & Winarske 1988; Tagliarini & Page 1988, pp. 775–82, Maa *et al* 1990, pp. 482–5). There are many situations where the problem can be formulated as minimization or maximization of a cost function or object function subjected to certain constraints. It is possible to map such a problem onto a feedback network, where the units and connection strengths are identified by comparing the cost function of the problem with the energy function of the network expressed in terms of the unit state values and the strengths of the connections. The solution to the problem lies in determining the state of the network at the global minimum of the energy function. In this process it is necessary to overcome the local minima of the energy function. This is accomplished by adopting a simulated annealing schedule for implementing the search for global minimum.

Probably the most studied problem in the context of optimization using principles of neural networks is the travelling salesman problem, where the objective is to find the shortest route connecting all cities to be visited by a salesman. Other optimization problems that are attempted include the weighted matching problem, where a number of points must be pairwise connected such that the sum of lengths of all connections is as short as possible, and stereo vision matching in optical image processing (Hertz *et al* 1991). The method of simulated annealing has also been successfully employed to find the optimal arrangement of integrated electronic circuits on semiconductor chips (Kirkpatrick *et al* 1983).

**6.1c Vector quantization:** Vector quantization (VQ) typically encodes a large set of training data vectors into a small set of representative points, thus achieving a significant compression in the representation of data. Vector quantization has been shown to be useful in compressing data that arises in image processing, speech processing, facsimile transmission, and weather satellites (Kohonen 1988; Nasrabadi & King 1988; Naylor & Li 1988).

Formally, vector quantization maps arbitrary data vectors to a binary representation or a symbol. The mapping is from an  $N$ -dimensional vector space to a finite set of symbols  $M$ . Associated with each symbol  $m \in M$  is a reproduction vector  $\hat{x}_m$ . The encoding of the data vector  $x$  to the symbol  $m$  is the mapping in VQ. The collection of all possible reproduction vectors is called the codebook.

The design of a codebook is called training, and it can be implemented using neural network models. The learning vector quantization (LVQ) structure is one such network model. Several other models have been proposed, for example, Kohonen's self-organising feature maps, to construct VQ codebooks for speech applications, and for image coding (Kohonen 1989; Ahalt *et al* 1990).

**6.1d Pattern classification:** Pattern classification is the most direct among all applications of neural networks. In fact neural networks became very popular because

of the ability of a multilayer feedforward neural network to form complex decision regions in the pattern space for classification (Gorman & Sejnowski 1988; Lippmann 1989). Many pattern recognition problems, especially character or other symbol recognition and vowel recognition, have been successfully implemented using multilayered networks (LeCun *et al* 1989; Pal & Mitra 1992). Note however that these networks are not directly applicable for situations where the patterns are deformed or modified due to transformations such as translation, rotation and scale change (Dotsenko 1988; Seibert & Waxman 1989).

## 6.2 Application areas

Neural network concepts and principles appear to have great potential for solving problems arising in practice. For most practical problems the solution by neural networks is not obvious. This is because the problems cannot be mapped directly onto an existing neural network architecture. In fact there are no principles guiding us to this mapping. There are many pattern recognition tasks in speech and vision which we seem to perform effortlessly, but we do not understand how we do so. For example, in speech, our auditory mechanism processes the signal directly in a manner suitable for later neural processing. On the other hand, to prepare input to an artificial neural network, the speech signal is normally processed in fixed frames of 10–20 ms to extract a fixed number of spectral or related parameters. In this process the temporal and spectral features with proper resolution needed for recognition may not have been captured. Moreover, there is as yet no neural network architecture which could perform the speech pattern recognition with the same effectiveness as human beings do. Similar comments apply to problems in visual pattern recognition also. Some of the other areas where human performance cannot be matched by existing neural architectures are in motor control and decision making.

Despite realization of these issues, there are several cases where neural principles have been used successfully. A few of them are listed below in different areas for illustration (Lisboa 1992).

**6.2a Speech processing:** Recognition of isolated utterances of characters in a speaker-independent mode over a telephone line has been demonstrated for directory enquiring application (Lang *et al* 1990; Cole *et al* 1992).

Medium-size (about 50 words) vocabulary speaker independent isolated word recognition using a partially connected network has been demonstrated to give equal or better performance compared to the conventional methods based on dynamic time warping (Bottou *et al* 1990).

Reliable discrimination of some stop consonants was demonstrated using time delay neural network architectures, and these ideas were extended to derive network architectures for syllable recognition (Waibel 1989).

Text-to-speech conversion with limited capabilities for English was demonstrated using multilayered feedforward neural networks (Sejnowski & Rosenberg 1987).

**6.2b Computer vision:** Recognition of hand-printed digits has been one of the most successful applications of neural networks (Krzyzak *et al* 1990). Satellite image data compression and enhancement of noisy images are some of the other useful applications (Hertz *et al* 1991; Raghu *et al* 1993).

Transformation invariant object recognition is one of the most difficult tasks,



although some impressive demonstration of neural network architectures are available for handwritten characters (Fukushima & Miyake 1982).

6.2c *Robotics and control*: Artificial vision for autonomous navigation, path planning with obstacle avoidance, and parallel computation of inverse dynamics are some of the applications of neural networks in robotics (Kung & Hwang 1989; Handleman *et al* 1990; Kuperstein & Wang 1990).

Operation guidance in blast furnace control and modelling nonlinearities in chemical process control are some of the applications of neural networks in control areas (Bhat & McAvoy 1989; Konishi *et al* 1990).

6.2d *Automated inspection and monitoring*: Explosive-detection in aircraft luggage, industrial quality control through visual inspection, forecasting for the utility industries, sonar signal identification and fault diagnosis for sensor failure in industrial plants are examples of the application of neural networks in inspection and monitoring situations (Shea & Lin 1989; Naidu *et al* 1990).

6.2e *Medical applications*: Medical diagnosis, noise filters for cardiac signals, image processing of ultrasonograms, and discrimination of signals for patient monitoring, have all been successfully implemented using networks (Reggia & Sutton 1988; Scalia *et al* 1988).

6.2f *Business and finance*: Scheduling and inventory control application, bond rating and asset forecasting in the stock market, exchange-rate forecasting, credit scoring, and mortgage underwriting have all demonstrated the successful use of neural network principles in business and finance (Collins *et al* 1988; Dutta & Shekhar 1988; White 1988).

## 7. Summary and Trends

In this tutorial article we have discussed the need for exploring new computing models for pattern recognition tasks. The importance of distinction between pattern processing and data processing has been discussed. The promise of the architectures inspired by the functions of biological neural networks has been shown by tracing the significant developments in artificial neural networks over the past decade. We have discussed the basics of artificial neural networks in terms of models of neurons, learning laws, and topology. We have also discussed the types of pattern recognition problems that can be solved by simple architectures based on the principles of artificial neural networks. Complex pattern recognition tasks require specialized architectures. Some general architectures were discussed for tasks requiring to resolve stability-plasticity dilemma and for tasks involving pattern variability and temporal patterns.

The most important issue for solving practical problems using the principles of ANN is still in evolving a suitable architecture to solve a problem. This continues to dominate this research area. ANN research may have to expand its scope to take into account the fuzzy nature of real world data and reasoning, and the complex (unknown) processing performed by the human perceptual mechanism through biological neural networks.

It is possible to view research in ANN along the following directions:

- (i) *Problem level*: Involves issues in mapping the real world problems as pattern processors. This may require good understanding of human information processing both from the psychological and the biological angle.
- (ii) *Basic level*: It is necessary to evolve better models of neurons as processing elements, their interconnections, dynamics (activation and synaptic), learning laws and recall procedures.
- (iii) *Functional level*: Involves development of basic structures which can solve a class of pattern recognition problems. These form building blocks for development of new architectures.
- (iv) *Architecture level*: This requires ideas to evolve new architectures from known principles, components and structures to solve complex pattern recognition problems. It is possible that the problems may be tailored somewhat to suit the architectures.
- (v) *Application level*: The objective is to solve a given practical problem using generally the principles of ANN but with ideas from other areas also like physics, signal processing etc.

This paper is mostly a consolidation of work reported by several researchers in the literature, some of which is cited in the references. The author has borrowed several ideas and illustrations from the references quoted in this paper.

The author would like to thank Mr M Babu for his assistance in preparing this paper and Dr H M Chouhan for his critical comments. The author also thanks the members of the Speech and Vision Laboratory for their interaction in the seminars on topics related to neural networks. Finally, this paper would not have come to this stage but for the initiative and interest shown by Prof. N Viswanadham of the Indian Institute of Science, Bangalore. The author is grateful to him for his encouragement.

## References

- Abu-Mostafa Y S, St. Jaques J M 1985 Information capacity of the Hopfield model. *IEEE Trans. Inf. Theor.* 31: 461-464
- Ackley D M, Hinton G E, Sejnowski T J 1985 A learning algorithm for Boltzmann machines. *Cogn. Sci.* 9: 147-169
- Ahalt S C, Krishnamurthy A K, Chen P, Melton D E 1990 Competitive learning algorithms for vector quantization. *Neural Networks* 3: 277-290
- Aleksander I, Morton H 1990 *An introduction to neural computing* (London: Chapman and Hall)
- Bhat N, McAvoy T 1989 Use of neural nets for dynamic modelling and control of chemical process systems. *Proc. Am. Autom. Contr. Conf.*, Pittsburgh, PA, pp. 1342-1348
- Bienenstock E, von der Malsburg Ch 1987 A neural network for the retrieval of superimposed connection patterns. *Euro. Phys. Lett.* 3: 1243-1249
- Bottou L, Soulie F F, Blanchet P, Lienard J S 1990 Speaker independent isolated digit recognition: multilayer perceptrons vs. dynamic time warping. *Neural Networks* 3: 436-465
- Carpenter G A 1989 Neural network models for pattern recognition and associative memory. *Neural networks* 2: 138-152
- Carpenter G A, Grossberg S 1987 ART2: Self-organization of stable category recognition codes for analog input patterns. *Appl. Opt.* 26: 4919-4930
- Carpenter G A, Grossberg S 1988 The ART of adaptive pattern recognition by a self-organizing neural network. *IEEE Comput.* 21: 77-88

- Cohen M, Grossberg S 1983 Absolute stability of global pattern formation and parallel storage by competitive neural networks. *IEEE Trans. Syst., Man Cybern.* SMC-13: 815-825
- Cole R, Fanty M, Muthuswamy Y, Gopalakrishna M 1992 Speaker-independent recognition of spoken English letters. *Proc. Int. Joint Conf. Neural Networks*, San Diego, CA
- Collins E, Ghosh S, Scotfield C L 1988 An application of a multiple neural network learning system to emulation of mortgage underwriting judgements. *IEEE Int. Conf. Neural Networks* (Piscataway, NJ: IEEE Press) 2: 459-466
- Cybenko G 1989 Continuous value neural networks with two hidden layers are sufficient. *Math. Control. Signal Syst.* 2: 303-314
- Desai M S 1990 *Noisy pattern retrieval using associative memories*. MSEE thesis, University of Louisville, Kentucky
- Deuker J, Schwartz D, Wittner B, Solla S, Howard R, Jackel L, Hopfield J 1987 Large automatic learning, rule extraction, and generalization. *Complex Syst.* 1: 877-922
- Dotsenko V S 1988 Neural networks: translation-, rotation- and scale invariant pattern recognition *J. Phys.* A21: L783-L787
- Dutta S, Shekhar S 1988 Bond rating: a non-conservative application of neural networks. *IEEE Int. Conf. Neural Networks* (Piscataway, NJ: IEEE Press) 2: 443-450
- Freeman J A, Skupura D M 1991 *Neural network algorithms, applications and programming techniques* (New York: Addison-Wesley)
- Fukushima K 1975 Cognitron: A self-organizing multilayer neural network. *Biol. Cybern.* 20: 121-136
- Fukushima K 1988 A neural network for visual pattern recognition. *IEEE Comput.* 21: 65-75
- Fukushima K 1991 Handwritten alphanumeric character recognition by the neocognitron. *IEEE Trans. Neural Networks* 2: 355-365
- Fukushima K, Miyake S 1982 Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recogn.* 15: 455-469
- Geman S, Geman D 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-6: 721-741
- Gorman R, Sejnowski T 1988 Learned classification of sonar targets using a massively parallel network. *IEEE Trans. Acoust. Speech Signal Process.* 36: 1135-1140
- Grossberg S 1969 Some networks that can learn, and reproduce any number of complicated space-time patterns. *Int. J. Math. Mech.* 19: 53-91
- Grossberg S 1980 How does a brain build a cognitive code? *Psychol. Rev.* 87: 1-51
- Grossberg S 1982 *Studies of mind & brain* (Boston: Reidel)
- Grossberg S 1988 Nonlinear neural networks: Principles, mechanisms, and architecture. *Neural networks* 1: 17-61
- Handleman D H, Lane S H, Gelfand J J 1990 Integrating neural networks and knowledge-based systems for intelligent robotic control. *IEEE Control Syst. Mag.* 10(3): 77-87
- Hassoun M H 1989 Dynamic heteroassociative memories. *Neural Networks* 2: 275-287
- Hebb D 1949 *Organization of the behaviour* (New York: Wiley)
- Hecht-Nielsen R 1987 Counterpropagation networks. *Appl. Opt.* 26: 4979-4984
- Hecht-Nielsen R 1990 *Neurocomputing* (Reading, MA: Addison-Wesley)
- Hertz J, Krogh A, Richard G P 1991 *Introduction to the theory of neural computation* (New York: Addison-Wesley)
- Hinton G E, Sejnowski T J 1986 Learning and relearning in Boltzmann machines. In *Parallel distributed processing: Explorations in the microstructure of cognition* (eds) D E Rumelhart, J L McClelland (Cambridge, MA: MIT Press) 1: 282-317
- Hodgkin A L, Huxley A F 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117: 500-544
- Hopfield J J 1982 Neural networks and physical systems with emergent collective computational capabilities. *Proc. Natl. Acad. Sci. (USA)* 79: 2554-2558
- Hopfield J J, Tank D W 1985 Neural computation of decisions in optimization problems. *Biol. Cybern.* 52: 141-154
- Huang Z, Kuh A 1992 A combined self-organizing feature map and multilayer perceptron for isolated word recognition. *IEEE Trans. Signal Process.* 40: 2651-2657
- Hush D R, Horne B G 1993 Progress in supervised neural networks. *IEEE Signal Process. Mag.* 10: 8-39
- Kamp Y, Hasler M 1990 *Recursive neural networks for associative memory* (Chichester: John Wiley & Sons)

- Kennedy M P, Chau L O 1988 Neural networks for nonlinear programming. *IEEE Trans. Circuits Syst.* CAS-35: 554-562
- Kirkpatrick S, Gelatt C D Jr, Vecchi M P 1983 Optimization by simulated annealing. *Science* 220: 671-680
- Kohonen T 1988 An introduction to neural computing. *Neural Networks* 1: 3-16
- Kohonen T 1989 *Self-organization and associative memory* (3rd edn) (Berlin: Springer-Verlag)
- Kohonen T 1990 The self-organizing map. *Proc. IEEE* 78: 1464-1480
- Konishi M, Otsuka Y, Matsuda K, Tamura N, Fuki A, Kadoguchi K 1990 Application of a neural network to operation guidance in a blast furnace. *3rd European Seminar on Neural Computing: The Marketplace*, London
- Kosko B 1988 Bidirectional associative memories. *IEEE Trans. Syst., Man Cybern.* 18: 49-60
- Kosko B 1990 Unsupervised learning in noise. *IEEE Trans. Neural Networks* 1: 44-57
- Kosko B 1992 *Neural networks and fuzzy systems* (Englewood Cliffs, NJ: Prentice-Hall)
- Krzyzak A, Dali W, Yuen C Y 1990 Unconstrained handwritten character classification using modified back propagation model. In *Frontiers in handwriting recognition* (ed.) C Y Suen (Montreal: CENPARMI)
- Kung S Y, Hwang J N 1989 Neural network architectures for robotic applications. *IEEE Trans. Robotics Autom.* 5: 641-657
- Kuperstein M, Wang J 1990 Neural controller for adaptive movements with unforeseen payloads. *IEEE Trans. Neural Networks* 1(1): 137-142
- Lang K J, Waibel A H, Hinton G E 1990 A time-delay neural network architecture for isolated word recognition. *Neural Networks* 3(1): 23-44
- LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D 1989 Back propagation applied to handwritten zip code recognition. *Neural Comput.* 1: 541-551
- Lippmann R P 1987 An introduction to computing with neural nets. *IEEE Trans. Acoust. Speech Signal Process. Mag.* (April): 4-22
- Lippmann R P 1989a Review of neural networks for speech recognition. *Neural Comput.* 1(1): 1-38
- Lippmann R P 1989b Pattern classification using neural networks. *IEEE Commun. Mag.* (Nov): 47-64
- Lisboa P G P 1992 *Neural networks current applications* (London: Chapman & Hall)
- Maa C Y, Chin C, Shanblatt M A 1990 A constrained optimization neural net techniques for economic power dispatch. *Proc. 1990* (New York: IEEE Press)
- Marcus A, van Dam A 1991 User-interface developments for the nineties. *IEEE Comput.* 24: 49-57
- McCulloch W S, Pitts W 1943 A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5: 115-133
- Michel A N, Farrell J A 1990 Associative memories via artificial neural networks. *IEEE Control Syst. Mag.* (April): 6-17
- Minsky M, Papert S A 1988 *Perceptron* (Cambridge, MA: MIT Press)
- Muller B, Reinhardt J 1990 *Neural networks: An introduction* (Berlin: Springer-Verlag)
- Murakami K, Aibara T 1987 An improvement on the Moore-Penrose generalized inverse associative memory. *IEEE Trans. Syst. Man. Cybern.* SMC 17: 699-706
- Naidu S R, Zafiriou E, McAvoy T J 1990 Use of neural networks for sensor failure detection in a control system. *IEEE Control Syst. Mag.* 10(3): 49-55
- Nasrabadi N M, King R A 1988 Image coding using vector quantization: A review. *IEEE Trans. Commun.* 36: 957-971
- Naylor J, Li K P 1988 Analysis of a neural network algorithm for vector quantization of speech parameters. *Neural Networks* 1 (Suppl): 310
- Pal S K, Mitra S 1992 Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Networks* 3: 683-697
- Raghu P P, Chouhan H M, Yegnanarayana B 1993 Multispectral image classification using neural network. *Proc. Natl. Conf. on Neural Networks*, Anna University, Madras (NCNN): 1-10
- Rauch H E, Winarske T 1988 Neural networks for routing communications traffic. *IEEE Control Syst. Mag.* (April): 26-31
- Reggia J A, Sutton G G 1988 III. Self-processing networks and their biomedical implications. *Proc. IEEE* 76: 680-692

- Rosenblatt F 1958 A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65: 386-408
- Rosenblatt F 1962 *Principles of neurodynamics* (Washington, DC: Spartan)
- Rumelhart D, McClelland J 1986 *Parallel distributed processing: Explorations in the microstructure of cognition* (Boston: MIT Press) vol. 1
- Rumelhart D E, Zipser D 1986 Feature discovery by competitive learning. *Parallel and distributed processing* (eds) J L McClelland, D E Rumelhart 1: 151-193
- Scalia F, Marconi L, Ridella S, Arrigo P, Mansi C, Mela G S 1988 An example of back propagation: diagnosis of dyspepsia. *Ist IEEE Conf. Neural Networks* (IEEE Conf. Publ.) 313: 332-540
- Schalkoft R 1992 *Pattern recognition - Statistical, structural and neural approaches* (New York: John Wiley & Sons)
- Seibert M, Waxman A 1989 Spreading activation layers, visual saccades, and invariant representations for neural pattern recognition systems. *Neural Networks* 2: 9-27
- Sejnowski T, Rosenberg C 1987 Parallel networks that learn to pronounce English text. *Complex Syst.* 1: 145-168
- Shea P M, Lin V 1989 Detection of explosives in checked airline baggage using an artificial neural system. *Int. Joint. Conf. on Neural Networks* 2: 31-34
- Simpson K P 1990 *Artificial neural systems* (New York: Pergamon)
- Simpson K P 1992 *Foundations of neural networks in artificial neural networks* (eds) Edgar Sanchez-Sinencio, Clifford Lau (New York: IEEE Press)
- Szu H 1986 Fast simulated annealing. In *Neural networks for computing* (ed.) J S Denker (New York: Snowbird)
- Tagliarini G A, Page E W 1988 A neural network solution to the concentrator assignment problem. *Neural information processing systems* (ed.) D Z Anderson (New York: Am. Inst. Phys.)
- von der Malsburg Ch 1973 Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* 14: 85-100
- Waibel A 1989 Modular construction of time-delay neural networks for speech recognition. *Neural Comput.* 1: 39-46
- Wasserman P D 1988 Combined backpropagation/cauchy machine. *Neural networks: Abstracts of the first INNS Meeting*, Boston (Elmsford, NY: Pergamon) 1: 556
- White H 1988 Economic prediction using neural networks: the case of IBM daily stock returns. *Neural networks: Abstracts of the First INNS Meeting*, Boston (Elmsford, NY: Pergamon) 1: 451-458
- Widrow B, Hoff M E 1960 Adaptive switching circuits. *IRE WESCON Convention Record* (4): 96-104
- Willshaw D J, von der Malsburg Ch 1976 How patterned neural connections can be set up by self-organization. *Proc. R. Soc. London B* 194: 431-445
- Zurada J M 1992 *Introduction to artificial neural systems* (St. Paul, MN: West)

## Convergence of higher-order two-state neural networks with modified updating

M VIDYASAGAR

Centre for AI & Robotics, Raj Bhavan Circle, High Grounds, Bangalore  
560 001, India

E-mail: [sagar@yantra.ernet.in](mailto:sagar@yantra.ernet.in)

MS received 9 September 1993; revised 25 February 1994

**Abstract.** The Hopfield network is a standard tool for maximizing a *quadratic* objective function over the discrete set  $\{-1, 1\}^n$ . It is well-known that if a Hopfield network is operated in an *asynchronous* mode, then the state vector of the network converges to a local maximum of the objective function; if the network is operated in a *synchronous* mode, then the state vector either converges to a local maximum, or else goes into a limit cycle of length two. In this paper, we examine the behaviour of *higher-order* neural networks, that is, networks used for maximizing objective functions that are not necessarily quadratic. It is shown that one can assume, without loss of generality, that the objective function to be maximized is *multilinear*. Three methods are given for updating the state vector of the neural network, called the asynchronous, the best neighbour and the gradient rules, respectively. For Hopfield networks with a quadratic objective function, the asynchronous rule proposed here reduces to the standard asynchronous updating, while the gradient rule reduces to synchronous updating; the best neighbour rule does not appear to have been considered previously. It is shown that both the asynchronous updating rule and the best neighbour rule converge to a local maximum of the objective function within a finite number of time steps. Moreover, under certain conditions, under the best neighbour rule, each global maximum has a nonzero radius of direct attraction; in general, this may not be true of the asynchronous rule. However, the behaviour of the gradient updating rule is not well-understood. For this purpose, a *modified* gradient updating rule is presented, that incorporates both *temporal as well as spatial* correlations among the neurons. For the modified updating rule, it is shown that, after a finite number of time steps, the network state vector goes into a limit cycle of length  $m$ , where  $m$  is the degree of the objective function. If  $m = 2$ , i.e., for quadratic objective functions, the modified updating rule reduces to the synchronous updating rule for Hopfield networks. Hence the results presented here are “true” generalizations of previously known results.

**Keywords.** Neural dynamics; Hopfield networks; higher-order networks; modified updating.

## 1. Introduction

A vast majority of the current research into feedback neural networks has been focused on the so-called *Hopfield networks*, whose dynamics are described by the equation

$$x_i(t+1) = \text{sign} \left[ \sum_{j=1}^n w_{ij} x_j(t) + \theta_i \right], \quad i = 1, \dots, n, \quad (1)$$

where  $n$  is the number of neurons,  $x_i(t) \in \{-1, 1\}$  is the state of neuron  $i$  at time  $t$ ,  $w_{ij}$  is the weight of the interconnection from neuron  $i$  to neuron  $j$ , and  $-\theta_i$  is the firing threshold of neuron  $i$ . Hopfield (1982) defines the energy of the network as<sup>1</sup>

$$E = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} x_i x_j + \sum_{i=1}^n x_i \theta_i, \quad (2)$$

and proves the following property: Suppose  $w_{ji} = w_{ij}$  for all  $i, j$  (symmetric interactions), and  $w_{ii} = 0$  for all  $i$  (no self-interactions). Finally, suppose the neural states are updated *asynchronously*, as follows: At each (discrete) instant of time  $t$ , select an integer  $i \in \{1, \dots, n\}$  at random, compute  $x_i(t+1)$  in accordance with (1), but leave  $x_j(t)$  unchanged for all  $j \neq i$ . In this mode of operation, it is true that

$$E[\mathbf{x}(t+1)] \geq E[\mathbf{x}(t)], \quad (3)$$

where  $\mathbf{x} = [x_1 \dots x_n]^t$ . Thus, in an asynchronous mode of operation, the neural network will eventually reach a fixed point of the network, that is, a vector  $\mathbf{x}_0$  with the property that

$$\mathbf{x}(t) = \mathbf{x}_0 \rightarrow \mathbf{x}(t+1) = \mathbf{x}_0, \quad (4)$$

irrespective of which neuron is updated at time  $t$ . Goles *et al* (1985) prove that, if the network is updated *synchronously*, that is, at time  $t$  the states of *all* neurons are updated according to (1), then the network either converges to a fixed point, or else goes into a limit cycle of length two. See Bruck & Goodman (1988) for a unification of both convergence results, and Kamp & Hasler (1990) for a book-length treatment of the topic.

Therefore, in the case where it is desired to optimize a quadratic function of the form (2) over a finite set  $\{-1, 1\}^n$ , the behaviour of the corresponding neural network (1) is well-understood. It is now known (see, e.g., Hopfield & Tank 1985) that several NP-complete problems can be formulated as the minimization of a quadratic function of the type (2). However, there are situations in which it is more natural to use an objective function which is a polynomial of degree three or higher. One such example is given in Masti & Vidyasagar (1991), wherein the problem of checking whether or not there exists a truth assignment on a set of Boolean variables that makes each of a set of formulas true (commonly known as the "satisfiability problem" and the "original" NP-complete problem), is formulated as a minimization problem over the set  $\{0, 1\}^n$  where  $n$  is the number of literals, and the degree of the objective function

<sup>1</sup>Actually, the energy function defined by Hopfield is the *negative* of this  $E$ , but this is a minor difference.

is the length of the longest clause in the set of formulas. Another example is given in Bruck & Blaum (1989), wherein the problem of algebraic block-coding is formulated as that of maximizing a polynomial over  $\{-1, 1\}^n$ , where the number of neurons  $n$  equals the length of the encoded words, and the degree of the objective function is equal to the number of information bits.

In this paper, we examine the use of neural networks for maximizing objective functions that are not necessarily quadratic, over the discrete set  $\{-1, 1\}^n$ . It is shown that one can assume, without loss of generality, that the objective function to be maximized is *multilinear*. Three methods are given for updating the state vector of the neural network, called the asynchronous, the best neighbour and the gradient rules, respectively. For Hopfield networks with a quadratic objective function, the asynchronous rule proposed here reduces to the standard asynchronous updating, while the gradient rule reduces to synchronous updating; the best neighbour rule does not appear to have been considered previously. It is shown that both the asynchronous updating rule and the best neighbour rule converge to a local maximum of the objective function within a finite number of time steps. Moreover, under certain conditions, under the best neighbour rule, each global maximum has a nonzero radius of direct attraction; in general, this may not be true of the asynchronous rule. However, the behaviour of the gradient updating rule is not well-understood. For this purpose, a *modified* gradient updating rule is presented, that incorporates both *temporal as well as spatial* correlations among the neurons. For the modified updating rule, it is shown that, after a finite number of time steps, the network state vector goes into a limit cycle of length  $m$ , where  $m$  is the degree of the objective function. If  $m = 2$ , i.e., for quadratic objective functions, the modified updating rule reduces to the synchronous updating rule for Hopfield networks. Hence the results presented here are "true" generalizations of previously known results.

## 2. Updating rules, fixed points and local maxima

In this section, we briefly study the problem of maximizing a polynomial  $E(\mathbf{x})$  as  $\mathbf{x}$  varies over the discrete set  $\{-1, 1\}^n$ . Three types of updating rules are introduced, namely: asynchronous, best neighbour, and gradient. The relationship between fixed points (under each type of updating rule) and local maxima is explored.

(1) *Asynchronous updating.* If  $\mathbf{x} \in \{-1, 1\}^n$  is the current state, choose an index  $i \in \{1, \dots, n\}$  at random. Define  $\mathbf{y} \in \{-1, 1\}^n$  by

$$y_i = -x_i, \quad y_j = x_j, \quad \text{for } j \neq i. \quad (5)$$

Note that  $\mathbf{y}$  is at a Hamming distance of one from  $\mathbf{x}$ . Set the next state equal to  $\mathbf{y}$  if  $E(\mathbf{y}) > E(\mathbf{x})$ , and equal to  $\mathbf{x}$  if  $E(\mathbf{y}) \leq E(\mathbf{x})$ . Note that if the next state is set equal to  $\mathbf{y}$  with a probability

$$p = 1/[1 + \exp(-\Delta E/T)], \quad (6)$$

where  $\Delta E = E(\mathbf{y}) - E(\mathbf{x})$  and  $T$  is the "temperature" of the network, then we get the simulated annealing algorithm (see e.g. van Laarhoven & Aarts 1987).

(2) *Best neighbour updating.* Suppose  $\mathbf{x} \in \{-1, 1\}^n$  is the current state. Let  $N(\mathbf{x})$



denote the set of nearest neighbours of  $\mathbf{x}$  in the sense of the Hamming distance; that is, let  $N(\mathbf{x})$  consist of all vectors in  $\{-1, 1\}^n$  that differ from  $\mathbf{x}$  in exactly one component. If  $E(\mathbf{y}) \leq E(\mathbf{x})$  for all  $\mathbf{y} \in N(\mathbf{x})$ , then set the next state equal to  $\mathbf{x}$ . Otherwise, set the next state to be a "best neighbour" of  $\mathbf{x}$ , that is, a vector  $\mathbf{y}_0 \in N(\mathbf{x})$  such that

$$E(\mathbf{y}_0) \geq E(\mathbf{y}), \quad \text{for all } \mathbf{y} \in N(\mathbf{x}). \quad (7)$$

The difference between asynchronous updating and best neighbour updating is that in the latter case,  $E$  is evaluated at *all* nearest neighbours of  $\mathbf{x}$  before the next state is determined, whereas in the former case,  $E$  is evaluated at a *single randomly chosen* neighbour of  $\mathbf{x}$ .

(3) *Gradient updating.* Let  $\mathbf{x} \in \{-1, 1\}^n$  be the current state. Let  $\nabla E(\mathbf{x}) \in \mathbb{R}^n$  denote the gradient of  $E$  evaluated at  $\mathbf{x}$ . Define the next state by

$$x_i(t+1) = \text{sign}[\nabla E(\mathbf{x})]_i, \quad i = 1, \dots, n, \quad (8)$$

or, more compactly,

$$\mathbf{x}(t+1) = \text{sign}[\nabla E(\mathbf{x})]. \quad (9)$$

If some component of  $\nabla E(\mathbf{x})$  equals zero, say  $[\nabla E(\mathbf{x})]_i = 0$ , then define  $x_i(t+1) = x_i(t)$ .

This brings up a subtle point that is worth stating explicitly. In both asynchronous and best neighbour updating rules, the expression for  $E(\mathbf{x})$  is unimportant, and only the set of values of  $E$  at the  $2^n$  points in  $\{-1, 1\}^n$  is important. However, in the gradient updating rule, the *form* of the function  $E(\mathbf{x})$  is also important. To illustrate this point, let  $E(\mathbf{x})$  be some function defined on  $\mathbb{R}^n$ , and define

$$E_1(\mathbf{x}) = x_1^2 E(\mathbf{x}). \quad (10)$$

Since  $x_1^2 = 1$  whenever  $x_1 = \pm 1$ , it follows that both  $E$  and  $E_1$  have the same values over  $\{-1, 1\}^n$ . Hence both the asynchronous and the best neighbour updating rules give identical updates irrespective of whether  $E$  or  $E_1$  is used as the objective function. However,

$$\frac{\partial E_1}{\partial x_1} = x_1^2 \frac{\partial E}{\partial x_1} + 2x_1 E = \frac{\partial E}{\partial x_1} + 2x_1 E \neq \frac{\partial E}{\partial x_1} \quad (11)$$

in general. Therefore, in the case of gradient updating, it *does* matter whether  $E$  or  $E_1$  is used as the objective function.

The next result shows a way around this difficulty by showing that, given the values of the objective function on  $\{-1, 1\}^n$ , there exists a *unique multilinear polynomial* that attains the same values. Recall that a function  $E(\mathbf{x})$  is said to be *multilinear* if, for each fixed index  $i \in \{1, \dots, n\}$ ,  $E(\mathbf{x})$  is an affine function of  $x_i$ . In other words, if  $E(\mathbf{x})$  is written out as a polynomial in the  $n$  variables  $x_1, \dots, x_n$ , no variable  $x_i$  appears with an exponent greater than one. This is a reasonable assumption, because  $x_i^2 = 1$  whenever  $\mathbf{x} \in \{-1, 1\}^n$ . If  $E(\mathbf{x})$  is a quadratic function of the form (2), then  $E$  is multilinear if and only if  $w_{ii} = 0$  for all  $i$ . More generally, a polynomial  $E$  is multilinear if and only if

$$\partial^2 E / \partial x_i^2 = 0, \quad \text{for all } i. \quad (12)$$

Thus, assuming that the objective function is a multilinear polynomial is a generalization of the "no self-interactions" assumption for Hopfield networks.

*Lemma 1.* Suppose  $E: \mathcal{R}^n \rightarrow \mathcal{R}$  is an arbitrary function. Then there exists a unique multilinear polynomial  $F: \mathcal{R}^n \rightarrow \mathcal{R}$  such that

$$E(\mathbf{z}) = F(\mathbf{z}), \quad \text{for all } \mathbf{z} \in \{-1, 1\}^n. \quad (13)$$

*Proof.* Given the polynomial  $E(\mathbf{z})$ , perform the substitution

$$z_i = 2x_i - 1, \quad x_i = (z_i + 1)/2. \quad (14)$$

Then it is easy to see that  $x_i = 1$  or  $0$ , if  $z_i = 1$  or  $-1$ . By making this substitution, the multilinear polynomial  $E(\mathbf{z})$  gets transformed into another polynomial  $\bar{E}(\mathbf{x})$ , which is also multilinear. Also, as  $\mathbf{z}$  varies over the  $2^n$  bipolar vectors in  $\{-1, 1\}^n$ , the corresponding vector  $\mathbf{x}$  varies over the  $2^n$  binary vectors  $\{0, 1\}^n$ . Hence, in order to prove the lemma, it is enough to establish that there exists a multilinear polynomial  $\bar{F}(\mathbf{x})$  such that

$$\bar{E}(\mathbf{x}) = \bar{F}(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \{0, 1\}^n. \quad (15)$$

Then, by back-substituting for  $\mathbf{x}$  in terms of  $\mathbf{z}$ , one can recover the desired multilinear polynomial  $F(\mathbf{z})$ . Hence attention is focused on establishing the relationship (15). For convenience, the "bars" on the symbols  $E$  and  $F$  are dropped.

Every multilinear polynomial in the variables  $x_1, \dots, x_n$  can be written in the form

$$F(\mathbf{x}) = \sum_{\mathbf{i} \in \{0, 1\}^n} c_{\mathbf{i}} x_1^{i_1} \cdots x_n^{i_n}, \quad (16)$$

where  $\mathbf{i} = \{i_1, \dots, i_n\}$  and  $0^0$  is taken as 1. Given the function  $E$ , determine the family of  $2^n$  (not necessarily distinct) values  $(E(\mathbf{x}), \mathbf{x} \in \{0, 1\}^n)$ . It is first shown that it is possible to select the coefficients  $c_{\mathbf{i}} \in \{0, 1\}^n$ , such that (15) holds. First, let  $\mathbf{x} = \mathbf{0}$ . Then, from (16),

$$F(\mathbf{0}) = c_{\mathbf{0}}. \quad (17)$$

Hence, set  $c_{\mathbf{0}} = E(\mathbf{0})$ . Next, let  $\mathbf{x} = \{0 \cdots 0 1 0 \cdots 0\}$ , where the 1 is in the  $j$ th position. Then

$$F(\mathbf{x}) = c_{\mathbf{0}} + c_{\{0 \cdots 0 1 0 \cdots 0\}}. \quad (18)$$

Thus set

$$c_{\{0 \cdots 0 1 0 \cdots 0\}} = E(\{0 \cdots 0 1 0 \cdots 0\}) - c_{\mathbf{0}}. \quad (19)$$

In this way, the  $n$  coefficients  $c_{\{1 0 \cdots 0\}}$  through  $c_{\{0 \cdots 0 1\}}$  can be determined. Next, choose  $\mathbf{x}$  to have exactly two nonzero components, and so on. At the end of the process, we will have chosen a set of coefficients  $c_{\mathbf{i}} \in \{0, 1\}^n$  such that the function  $F$  defined by (16) satisfies (15). To show that this choice is unique, suppose  $\bar{F}$  is another multilinear polynomial that also satisfies (15) (with  $F$  replaced by  $\bar{F}$  of course). Then  $\Delta F(\mathbf{x}) = F(\mathbf{x}) - \bar{F}(\mathbf{x})$  is also a multilinear polynomial, and  $\Delta F(\mathbf{x}) = 0$  for all  $\mathbf{x} \in \{0, 1\}^n$ . Now a repetition of the above argument shows that all coefficients of  $\Delta F$  are zero; that is,  $\Delta F(\mathbf{x}) = 0$  for all  $\mathbf{x} \in \mathcal{R}^n$ .  $\square$

*Example 1.* Suppose  $n = 2$ , and let

$$E(z_1, z_2) = z_1 \exp(z_2).$$

It is desired to find a multilinear polynomial  $F(z_1, z_2)$  such that (13) holds. For this purpose, we first make the transformation

$$z_1 = 2x_1 - 1, \quad z_2 = 2x_2 - 1.$$

This gives

$$\bar{E}(x_1, x_2) = (2x_1 - 1) \exp(2x_2 - 1).$$

Now the function  $\bar{E}$  is evaluated at the four vectors  $(x_1, x_2) = (0, 0), \dots, (1, 1)$ . This gives

$$\bar{E}(0, 0) = E(-1, -1) = -e^{-1}, \quad \bar{E}(1, 0) = E(1, -1) = e^{-1},$$

$$\bar{E}(0, 1) = E(-1, 1) = -e, \quad \bar{E}(1, 1) = E(1, 1) = e.$$

Now set up the multilinear form

$$\bar{F}(x_1, x_2) = c_{00} + c_{10}x_1 + c_{01}x_2 + c_{11}x_1x_2.$$

In order for (15) to hold, it is necessary that

$$c_{00} = \bar{E}(0, 0) \Rightarrow c_{00} = -e^{-1},$$

$$c_{00} + c_{10} = \bar{E}(1, 0) \Rightarrow c_{10} = 2e^{-1},$$

$$c_{00} + c_{01} = \bar{E}(0, 1) \Rightarrow c_{01} = -e + e^{-1},$$

$$c_{00} + c_{10} + c_{01} + c_{11} = \bar{E}(1, 1) \Rightarrow c_{11} = 2e - 2e^{-1}.$$

This uniquely determines  $\bar{F}(x_1, x_2)$ . The desired multilinear polynomial  $F(z_1, z_2)$  that satisfies (13) can now be determined by substituting

$$x_1 = (2z_1 + 1)/2, \quad x_2 = (2z_2 + 1)/2.$$

The details are omitted as they are rather simple.  $\square$

If the function  $E$  is already a polynomial, then it is routine to determine the corresponding multilinear polynomial  $F$  such that (13) holds. It is only necessary to replace each term of the form  $x_i^{2k}$  by 1, and each term of the form  $x_i^{2k+1}$  by  $x_i$ , because  $x_i^2 = 1$  whenever  $x_i \in \{-1, 1\}^n$ . For example, consider the polynomial  $E: \mathbb{R}^4 \rightarrow \mathbb{R}$  defined by

$$E(\mathbf{x}) = x_1^2 x_2 x_3 + x_1 x_2^3 x_4 + x_2 x_4 + x_3.$$

Then

$$F(\mathbf{x}) = x_2 x_3 + x_1 x_2 x_4 + x_2 x_4 + x_3.$$

## COROLLARY 2

Define the  $2^n \times 2^n$  matrix  $M$  as follows: Order the  $2^n$  elements of  $\{-1, 1\}^n$  in some order, and do the same for the  $2^n$  elements of  $\{0, 1\}^n$ . Let

$$m_{i,x} = x_1^{i_1} \cdots x_n^{i_n}, \quad i \in \{0, 1\}^n, \quad x \in \{-1, 1\}^n. \quad (20)$$

Then  $M$  is nonsingular.

With the aid of lemma 1, we can find a *unique* multilinear polynomial  $E$  with prescribed values on  $\{-1, 1\}^n$ . If  $E$  is chosen as this polynomial, then the gradient updating rule is unambiguous.

Next, the concepts of fixed point and local maximum are introduced.

#### DEFINITION 1

A vector  $x_0 \in \{-1, 1\}^n$  is said to be a *fixed point* of the neural network with respect to the objective function  $E$  and a particular updating rule, if

$$x(0) = x_0 \Rightarrow x(1) = x_0, \quad (21)$$

under that updating rule.

Note that, if the asynchronous updating rule is used, then a vector  $x_0$  must satisfy (21) irrespective of which index  $i$  is chosen, in order to qualify as a fixed point.

#### DEFINITION 2

A vector  $x \in \{-1, 1\}^n$  is said to be a *local maximum* of the objective function  $E$  if

$$E(x) \geq E(y), \quad \text{for all } y \in N(x), \quad (22)$$

where  $N(x)$  denotes the set of  $n$  vectors that lie at a Hamming distance of one from  $x$ . It is said to be a *strict local maximum* if

$$E(x) > E(y), \quad \text{for all } y \in N(x). \quad (23)$$

As a prelude to the main result of this section, we give an alternate interpretation of asynchronous updating.

*Lemma 3. Define an updating rule as follows: Given the current state  $x(t)$  and an objective function  $E(x)$ , choose an index  $i \in \{1, \dots, n\}$  at random, and define*

$$x_i(t+1) = \text{sign}\{\nabla E[x(t)]_i\}, \quad x_j(t+1) = x_j(t), \quad j \neq i. \quad (24)$$

*If the objective function  $E$  is multilinear, then the above updating rule is identical to the asynchronous updating rule.*

*Note.* The assumption that  $E$  is multilinear is essential; the lemma is *not* true for arbitrary objective functions.

*Proof.* Let  $\bar{x}_i$  denote the  $(n-1)$ -dimensional vector obtained by omitting the  $i$ th component of  $x$ ; that is,

$$\bar{x}_i = [x_1 \cdots x_{i-1} x_{i+1} \cdots x_n]^T. \quad (25)$$

The multilinearity of the function  $E$  implies that there exist functions  $\bar{E}_i: \mathcal{R}^{n-1} \rightarrow \mathcal{R}$  and  $c_i: \mathcal{R}^{n-1} \rightarrow \mathcal{R}$  such that

$$E(x) = x_i \bar{E}_i(\bar{x}_i) + c_i(\bar{x}_i). \quad (26)$$

Notice also that  $\bar{E}_i(\bar{x}_i) = [\nabla E(\mathbf{x})]_i$ . Therefore, if the  $i$ th component of  $\mathbf{x}(t)$  is replaced by  $-x_i(t)$ , then the resulting change in  $E$  is given by

$$\Delta E = -2x_i(t)\{\nabla E[\mathbf{x}(t)]_i\}. \quad (27)$$

In the asynchronous updating rule,  $x_i(t+1)$  is set equal to  $-x_i(t)$  if and only if  $\Delta E > 0$ . It is now clear that the proposed rule is identical to the asynchronous updating rule.  $\square$

Now we come to the main result of this section.

**Theorem 1.** Suppose  $E$  is a multilinear polynomial on  $\mathcal{R}^n$ . Suppose  $\mathbf{x} \in \{-1, 1\}^n$ , and that no component of  $\nabla E(\mathbf{x})$  is zero. Under these conditions, the following statements are equivalent:

1.  $\mathbf{x}$  is a strict local maximum of  $E$ .
2.  $\mathbf{x}$  is a fixed point under asynchronous updating.
3.  $\mathbf{x}$  is a fixed point under best neighbour updating.
4.  $\mathbf{x}$  is a fixed point under gradient updating.

*Proof.* Suppose  $\mathbf{y} \in N(\mathbf{x})$  is obtained by replacing  $x_i$  by  $-x_i$ . Let  $i$  be an arbitrary index from the set  $\{1, \dots, n\}$ . Define  $\bar{x}_i$ ,  $\bar{E}_i$  and  $c_i$  as in the proof of lemma 3. Then, from (27),

$$E(\mathbf{y}) - E(\mathbf{x}) = -2x_i\bar{E}_i(\bar{x}_i) = -2x_i[\nabla E(\mathbf{x})]_i.$$

Now  $\mathbf{x}$  is a strict local maximum of  $E$  if and only if  $E(\mathbf{y}) - E(\mathbf{x}) < 0$  for all  $\mathbf{y} \in N(\mathbf{x})$ , or equivalently

$$x_i = \text{sign}[\nabla E(\mathbf{x})]_i, \quad i = 1, \dots, n. \quad (29)$$

Clearly, (29) is also a necessary and sufficient condition for  $\mathbf{x}$  to be a fixed point under each of the three types of updating.  $\square$

At this point the reader may wonder why a distinction is made between asynchronous updating and best neighbour updating. The reason is brought out next.

Suppose  $\mathbf{x}_0$  is a fixed point under some updating rule. Then the domain of direct attraction of  $\mathbf{x}_0$ , denoted by  $\text{DDA}(\mathbf{x}_0)$ , is defined as

$$\text{DDA}(\mathbf{x}_0) = \{\mathbf{x} \in \{-1, 1\}^n : \mathbf{x}(0) = \mathbf{x} \Rightarrow \mathbf{x}(1) = \mathbf{x}_0\}. \quad (30)$$

In other words,  $\text{DDA}(\mathbf{x}_0)$  is the set of states that get mapped into  $\mathbf{x}_0$  in a single time step. The radius of direct attraction of  $\mathbf{x}_0$ , denoted by  $\text{RDA}(\mathbf{x}_0)$ , is defined as the largest number  $r$  such that

$$H(\mathbf{x}, \mathbf{x}_0) \leq r \Rightarrow \mathbf{x} \in \text{DDA}(\mathbf{x}_0), \quad (31)$$

where  $H(\mathbf{x}, \mathbf{x}_0)$  is the Hamming distance between  $\mathbf{x}$  and  $\mathbf{x}_0$ , that is, the number of components in which  $\mathbf{x}$  and  $\mathbf{x}_0$  differ. In other words, the radius of direct attraction is the radius of the largest "sphere" centered at  $\mathbf{x}_0$  contained in the set  $\text{DDA}(\mathbf{x}_0)$ .

**Theorem 2.** Suppose  $\mathbf{x}_0$  is a global maximum of the function  $E$ ; that is,  $E(\mathbf{x}_0) \geq E(\mathbf{y})$

for all  $y \in \{-1, 1\}^n$ . Suppose also that all other global maxima of  $E$  are at a distance of at least three from  $x_0$ . Then

1.  $x_0$  is a fixed point under best neighbour updating.
2.  $RDA(x_0) \geq 1$ .

*Proof.* The proof is almost obvious. The hypotheses on  $x_0$  imply that  $E(x_0) > E(x)$  for all  $x \in N(x_0)$ . Hence  $x_0$  is a fixed point under best neighbour updating. Next, suppose  $x \in N(x_0)$ , and that  $y \in N(x)$ . Then  $H(y, x_0) \leq 2$ . By assumption, this implies that  $E(x_0) > E(y)$  (unless  $y = x_0$ ), because there is no global maximum of  $E$  within a Hamming distance of 2 from  $x_0$ . Hence  $x_0$  is the best neighbour of  $x$ , which means that if  $x(0) = x$ , then  $x(1) = x_0$  under best neighbour updating.  $\square$

The only reason for mentioning this obvious result is that it is *not* true under asynchronous updating.

*Example 2.* As an illustration, consider the block-coding problem studied in Bruck & Blaum (1989). For the (7, 4) Hamming code, the integer  $n$  equals 7, and the objective function is

$$E(x) = x_1 x_2 x_4 x_5 + x_1 x_3 x_4 x_6 + x_1 x_2 x_3 x_7. \quad (32)$$

It is easy to see that  $E(x) \leq 3$  for all  $x \in \{-1, 1\}^7$ . Moreover,  $E(x) = 3$  if and only if

$$x_1 x_2 x_4 x_5 = x_1 x_3 x_4 x_6 = x_1 x_2 x_3 x_7 = 1. \quad (33)$$

There are exactly  $2^4 = 16$  vectors that satisfy (33), and these are the global maxima of  $E$ . Moreover, the minimum Hamming distance between any pair of global maxima is 3. For details, see Bruck & Blaum (1989). Thus, by theorem 2, each of these global maxima has a radius of direct attraction of at least 1, if best neighbour updating is used.

To show that this is not true if asynchronous updating is used, let  $x_0$  equal the vector of all 1's. It is clear that  $x_0$  is a global maximum of  $E$ , because (33) is satisfied. Now consider the sequence

$$x(0) = [-1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1], \quad (34)$$

$$x(1) = [-1 \ 1 \ 1 \ 1 \ -1 \ 1 \ 1], \quad (35)$$

$$x(2) = [-1 \ 1 \ 1 \ 1 \ -1 \ -1 \ 1], \quad (36)$$

$$x(3) = [-1 \ 1 \ 1 \ 1 \ -1 \ -1 \ -1]. \quad (37)$$

Then

$$E(0) = -3, \quad E(1) = -1, \quad E(2) = 1, \quad E(3) = 3. \quad (38)$$

This is a valid sequence under asynchronous updating, because the objective function is strictly increased at each time step. Moreover,  $x(3)$  is a global maximum of  $E$ , because  $E(3) = 3$ . Hence  $x(3)$  is a fixed point under asynchronous updating. Note that  $x(1) \in N(x_0)$ . Therefore, starting at a neighbour of  $x_0$ , we have constructed a sequence of transitions that converge to *another* fixed point. This shows that, under asynchronous updating, the radius of direct attraction is zero.

It can be shown that, with the objective function suggested in Bruck & Blaum (1989), the same phenomenon occurs with *every* choice of the coding matrix and *every* equilibrium.  $\square$

In view of theorem 1, the next result, which discusses the convergence of the neural network under asynchronous and best neighbour updating, is almost obvious.

**Theorem 3.** *Suppose the network is operated under the asynchronous updating rule or the best neighbour updating rule. Then the network converges to a local maximum in a finite number of time steps.*

### 3. Modified synchronous updating

Theorem 3 shows that, under both the asynchronous as well as the best-neighbour updating rules, the trajectory of the neural network (9) converges to a local maximum of the objective function within a finite number of time steps. But what happens if the network is updated according to the gradient rule? This corresponds to *synchronous* updating of all neurons according to (9). In the case of a Hopfield network of the form (1), it is known that the network trajectory either converges to a local maximum or else goes into a limit cycle of length two (see Goles *et al* (1985)). But in the case where the objective function is not necessarily quadratic, the behaviour of the network under gradient updating is not well-understood. In this section, it is shown that, by *modifying* the gradient updating formula, it is possible to say something about the convergence of the network.

The traditional Hopfield network described by (1) is characterized by *linear* interconnections; that is, the right side of (1) is an affine function of the neural state vector  $\mathbf{x}(t)$ . In case where a neural network is to be used to optimize a polynomial of degree three or more, some authors propose the use of higher-order interconnections among neurons. The gradient updating rule (9) is an example of higher-order interconnections, because if  $E$  is a polynomial of degree three or more, then each component of  $\nabla E(\mathbf{x})$  is a polynomial of degree two or more. Higher-order interconnections can be said to reflect *spatial* correlation between neural states. However, as mentioned in the preceding paragraph, the behaviour of a neural network under gradient updating is not well-understood. In this section, we propose a modification of gradient updating for higher-order neural networks. In the modified formula, the state of the network at time  $t+1$  depends on the state of the network at times  $t, t-1, \dots, t-(m-2)$ , where  $m$  is the degree of the objective function. If the objective function is quadratic, then  $m=2$  and  $m-2=0$ . In this case, the state of the network at time  $t+1$  depends only on its state at time  $t$ , and the updating formula reduces to (1). For higher-order objective functions, however, the updating formula makes use of the past states as well. Moreover, the interaction terms are products of up to  $m-1$  terms. Hence the updating formula makes use of *temporal as well as spatial* correlation between neurons. With this updating formula, it is shown that the network trajectory either converges to an equilibrium, or else goes into a limit cycle of length  $m$ , where  $m$  is the degree of the objective function. Hence theorem 5 below is a "true" generalization of the results in Goles *et al* (1985).

As a prelude to presenting the updating formula, we give a brief discussion on symmetric multilinear forms. Suppose we are given an objective function  $E$ , which is a multilinear polynomial of degree  $m$ . Then we can write

$$E(\mathbf{x}) = \sum_{k=1}^m E_k(\mathbf{x}, \dots, \mathbf{x}), \quad (39)$$

where  $E_k$  is a homogeneous, symmetric, multilinear polynomial. Homogeneity means simply that  $E_k$  comprises all terms in  $E$  that are a product of *exactly*  $k$  terms of the form  $x_{i_1}, \dots, x_{i_k}$ . Multilinearity in the present context means that  $E_k(x_1, \dots, x_k)$  is a linear function of  $x_i$  for each  $i \in \{1, \dots, k\}$ . Symmetry means that

$$E_k(x_1, \dots, x_k) = E_k(x_{\pi(1)}, \dots, x_{\pi(k)}), \quad (40)$$

for all permutations  $\pi$  of  $\{1, \dots, k\}$  into itself. For instance, if  $k = 2$ , then

$$E_2(x, y) = x^t M y \quad (41)$$

for some  $n \times n$  matrix  $M$ . Now  $E_2$  is symmetric if and only if the matrix  $M$  is symmetric.

If  $E_k$  is a multilinear homogeneous polynomial of degree  $k$ , it is easy to see that there exists a function  $e_k: \mathcal{R}^{n(k-1)} \rightarrow \mathcal{R}^n$  such that

$$E_k(x_1, \dots, x_k) = x_1^t e_k(x_2, \dots, x_k), \quad (42)$$

where each component of the vector  $e_k(x_2, \dots, x_k)$  is a homogeneous, multilinear polynomial of degree  $k - 1$ . Moreover, if  $E_k$  is symmetric, so is  $e_k$ , in the sense that

$$e_k(x_2, \dots, x_k) = e_k(x_{\pi(2)}, \dots, x_{\pi(k)}), \quad (43)$$

for all permutation  $\pi$  of  $\{2, \dots, k\}$  into itself.

For every homogeneous multilinear polynomial  $E_k$  of degree  $k$ , there is an equivalent homogeneous *symmetric* polynomial  $E_{sk}$  such that

$$E_{sk}(x, \dots, x) = E_k(x, \dots, x). \quad (44)$$

In fact, we can define

$$E_{sk}(x_1, \dots, x_k) = \frac{1}{k!} \sum_{\pi} E_k(x_{\pi(1)}, \dots, x_{\pi(k)}), \quad (45)$$

where the summation is over all permutations  $\pi$  of  $\{1, \dots, k\}$  into itself. For instance, if  $k = 2$  and

$$E_2(x, y) = x^t M y, \quad (46)$$

then

$$E_{s2}(x, y) = \frac{1}{2} [E(x, y) + E(y, x)] = \frac{1}{2} x^t [M + M^t] y, \quad (47)$$

which is the familiar method of obtaining a *symmetric* quadratic form. The procedure in the case of higher values of  $k$  is illustrated through an example.

*Example 1.* (Continued) Consider once again the objective function in connection with the (7, 4) Hamming code, namely

$$E(x) = x_1 x_2 x_4 x_5 + x_1 x_3 x_4 x_6 + x_1 x_2 x_3 x_7. \quad (48)$$

In this case,  $E(x)$  is a homogeneous polynomial of degree 4. If we define

$$E_4(x, y, z, v) = x_1 y_2 z_4 v_5 + x_1 y_3 z_4 v_6 + x_1 y_2 z_3 v_7, \quad (49)$$



then  $E_4$  is not symmetric. It can be replaced by an equivalent symmetric polynomial as follows: Because of the complexity, the details are given only for the first term, namely  $x_1 y_2 z_4 v_5$ . There are  $4! = 24$  ways of assigning subscripts 1, 2, 4, 5 to the symbols  $x, y, z, v$ . Write down all 24 possible combinations, add them up, and divide by 24. Thus

$$E_{s4}(x, y, z, v) = (1/24)[x_1 y_2 z_4 v_5 + x_2 y_1 z_4 v_5 + \cdots x_5 y_4 z_2 v_1]. \quad (50)$$

Symmetric versions of the other terms can be constructed in an entirely analogous manner.

Now we are ready to state the updating rule. Suppose the objective function  $E(x)$  is a multilinear polynomial of degree  $m$  in  $x_1, \dots, x_n$ . Then, without loss of generality, we can write

$$E(x) = \sum_{k=1}^m E_k(x, \dots, x), \quad (51)$$

where  $E_k$  is a homogeneous, symmetric, multilinear polynomial.<sup>2</sup> Define  $e_k$  as before, namely by

$$E_k(x_1, \dots, x_k) = x'_1 e_k(x_2, \dots, x_k), \quad (52)$$

where each component of  $e_k$  is a homogeneous, symmetric, multilinear polynomial of degree  $k-1$ . Note that

$$\nabla E_k(x) = k e_k(x, \dots, x). \quad (53)$$

The idea behind the modified updating rule is rather simple, though the notation can become somewhat cumbersome. To state it, it is helpful to introduce the symbol  $S(k, m)$ . Given integers  $m > k > 0$ , let  $S(k, m)$  denote the set of all  $k$ -tuples  $(i_1, \dots, i_k)$  such that

$$0 \leq i_1 < i_2 < \cdots < i_k \leq m-1. \quad (54)$$

Note that this is a minor modification of the symbol  $S(m, n)$  in Vidyasagar (1985, p. 391). For example,  $S(3, 5)$  equals

$$S(3, 5) = \{(0, 1, 2), (0, 1, 3), (0, 1, 4), (0, 2, 3), (0, 2, 4), (0, 3, 4), \\ (1, 2, 3), (1, 2, 4), (1, 3, 4), (2, 3, 4)\}. \quad (55)$$

It is easy to see that the number of elements in  $S(k, m)$  equals  $m!/k!(m-k)!$ .

Given an objective function of the form (51), we first define the constants  $\alpha_k$ ,  $k = 1, \dots, m$ , by

$$\alpha_k = [(m-k)!k!]/(m-1)!. \quad (56)$$

Note that  $\alpha_k = m/|S(k, m)|$ , where  $|S(k, m)|$  is the cardinality of the set  $S(k, m)$ , i.e., the number of elements in the set  $S(k, m)$ . For each  $k$  in  $\{1, \dots, m\}$ , define the function

<sup>2</sup>Note that we use  $E_k$  instead of  $E_{sk}$ , which would be consistent with earlier notation.

$f_k(t)$  as follows: Let  $\pi$  vary over all elements of the set  $S(k-1, m-1)$ . For each  $\pi$ , define

$$f_{k,\pi}(t) = e_k[x(t-\pi_1), \dots, x(t-\pi_{k-1})], \quad (57)$$

and define

$$f_k(t) = \sum_{\pi \in S(k-1, m-1)} f_{k,\pi}(t). \quad (58)$$

Now the updating rule is as follows:

$$x(t+1) = \text{sign} \left[ \sum_{k=1}^m \alpha_k f_k(t) \right]. \quad (59)$$

*Example 2.* Suppose  $E$  is of degree 4. Express  $E(x)$  as

$$E(x) = E_4(x, x, x, x) + E_3(x, x, x) + E_2(x, x) + E_1(x) + E_0. \quad (60)$$

Define the vectors  $e_4, e_3, e_2, e_1$  in accordance with (3.4). The sets  $S(k-1, 3)$  are given as

$$\begin{aligned} S(3, 3) &= \{(0, 1, 2)\}, \\ S(2, 3) &= \{(0, 1), (0, 2), (1, 2)\}, \\ S(1, 3) &= \{0, 1, 2\}. \end{aligned} \quad (61)$$

The set  $S(0, 3)$  is undefined (see below). The vectors  $f_4, f_3, f_2, f_1$  are given respectively by

$$\begin{aligned} f_4 &= e_4[x(t), x(t-1), x(t-2)], \\ f_3 &= e_3[x(t), x(t-1)] + e_3[x(t), x(t-2)] + e_3[x(t-1), x(t-2)], \\ f_2 &= e_2[x(t)] + e_2[x(t-1)] + e_2[x(t-2)], \\ f_1 &= e_1, \end{aligned} \quad (62)$$

where we take advantage of the fact that, because  $E_1(x)$  is a linear functional of  $x$ , the gradient  $e_1$  is independent of  $x$ . The constants  $\alpha_i$  are given from (56) as

$$\alpha_4 = 4, \quad \alpha_3 = 1, \quad \alpha_2 = 2/3, \quad \alpha_1 = 1. \quad (63)$$

Finally, the updating rule is given by (59).

It is left to the reader to verify that, if  $m=2$  so that the objective function is quadratic and has the form

$$E(x) = \frac{1}{2}x'Wx + x'\theta, \quad (64)$$

then the updating rule becomes

$$x(t+1) = \text{sign}[Wx(t) + \theta], \quad (65)$$

which is the familiar rule. Hence (59) is a "true" generalization of the Hopfield network.

What is the advantage gained by using the complicated updating rule (59) in place of the simpler gradient rule

$$x(t+1) = \text{sign}\{\nabla E[x(t)]\}? \quad (66)$$

The principal advantage is that the dynamics of the system (59) can be analysed precisely. This is the main result of this section, and indeed, of the paper. Before stating this result, a preliminary issue is resolved.

### DEFINITION 3.

A vector  $\mathbf{x}_0 \in \{-1, 1\}^n$  is said to be a *fixed point* of the system (59) if

$$\mathbf{x}(0) = \mathbf{x}(1) = \dots = \mathbf{x}(m-2) = \mathbf{x}_0 \Rightarrow \mathbf{x}(t) = \mathbf{x}_0 \text{ for all } t \geq m-1. \quad (67)$$

This is just the standard definition of a fixed point, adapted for the fact that the right side of (59) contains some delayed terms.

**Theorem 4.** *A vector  $\mathbf{x}_0 \in \{-1, 1\}^n$  is a fixed point of (59) if and only if it is a fixed point of the gradient updating rule, that is, if and only if*

$$\mathbf{x}_0 = \text{sign}[\nabla E(\mathbf{x}_0)]. \quad (68)$$

Theorem 4 means that, even though there are time delays present in the updating rule (59), the fixed points are the same as those of the gradient updating rule. Combined with theorem 1, this result means that, should the state vector of the network converge to a fixed point, this fixed point will be a local maximum of the objective function  $E$ , even though time delays are introduced into the updating rule.

*Proof.* Suppose  $\mathbf{x}_0$  satisfies (67), and suppose  $\mathbf{x}(0) = \mathbf{x}(1) = \dots = \mathbf{x}(m-2) = \mathbf{x}_0$ . Then it follows from (57) that

$$\mathbf{f}_{k,\pi} = \mathbf{e}_k(\mathbf{x}_0, \dots, \mathbf{x}_0). \quad (69)$$

Therefore, from (58),

$$\mathbf{f}_k = \frac{(m-1)!}{(k-1)!(m-k)!} \mathbf{e}_k(\mathbf{x}_0, \dots, \mathbf{x}_0), \quad (70)$$

where the coefficient on the right side is the number of elements of the set  $S(k-1, m-1)$ . Now combining (56) and (70) shows that

$$\alpha_k \mathbf{f}_k = k \mathbf{e}_k(\mathbf{x}_0, \dots, \mathbf{x}_0) = \nabla E_k(\mathbf{x}_0) \quad (71)$$

Finally,

$$\sum_{k=1}^m \alpha_k \mathbf{f}_k = \sum_{k=1}^m \nabla E_k(\mathbf{x}_0) = \nabla E(\mathbf{x}_0). \quad (72)$$

Therefore (67) is satisfied if and only if

$$\mathbf{x}_0 = \text{sign}[\nabla E(\mathbf{x}_0)], \quad (73)$$

that is, if and only if  $\mathbf{x}_0$  is a fixed point under the gradient updating rule.  $\square$

Now we come to the main result.

**Theorem 5.** *With the updating rule (59), there exists a finite integer  $N$  such that*

$$\mathbf{x}(t) = \mathbf{x}(t-m), \text{ for all } t \geq N. \quad (74)$$

**Remark.** Theorem 5 states that, with the modified updating rule, each trajectory of the neural network eventually satisfies (74). This means that the possible behaviours of the network are: (i) it settles into a fixed point, or (ii) it goes into a limit cycle whose length is *divisor* of  $m$ . For example, if the objective function has degree 6, the network trajectories either converge to a fixed point, or else go into a limit cycle whose length is 2, 3, or 6.

**Proof.** Define a function  $G(t)$  as follows:

$$G(t) = \sum_{k=1}^m \alpha_k G_k(t), \quad (75)$$

where  $\alpha_k$  is defined in (56), and  $G_k(t)$  is defined as follows: Let  $\pi$  vary over  $S(k, m)$ , and define

$$G_{k,\pi}(t) = E_k[\mathbf{x}(t - \pi_1), \dots, \mathbf{x}(t - \pi_k)], \quad (76)$$

$$G_k(t) = \sum_{\pi \in S(k, m)} G_{k,\pi}(t). \quad (77)$$

Now let us compute the quantity

$$\Delta G(t) = G(t+1) - G(t). \quad (78)$$

It is claimed that

$$\Delta G(t) = [\mathbf{x}(t+1) - \mathbf{x}(t-m+1)]^t \left[ \sum_{k=1}^m \alpha_k \mathbf{f}_k(t) \right], \quad (79)$$

where  $\mathbf{f}_k(t)$  is defined in (58). Suppose for the moment that (79) is true. Then, in view of (59), it follows that if  $\mathbf{x}(t+1) \neq \mathbf{x}(t-m+1)$ , then  $\Delta G(t) > 0$ . Moreover, the increase  $\Delta G(t)$  can be bounded from below. In fact,  $\Delta G(t)$  is at least equal to the magnitude of the smallest component of  $\sum_k \alpha_k \mathbf{f}_k(t)$ . Since  $\mathbf{x}(t)$  varies over a finite set, this quantity itself can be bounded from below, say by  $\varepsilon$ . Similarly,  $G(t)$  can be bounded from above, say by  $M$ . Thus it follows that  $\mathbf{x}(t+1) \neq \mathbf{x}(t-m+1)$  cannot happen more than  $[M/\varepsilon]$  times, that is, a finite number of times.

Thus the proof is complete if it can be shown that (79) holds. What we show instead is that

$$\Delta G_k(t) = G_k(t+1) - G_k(t) = [\mathbf{x}(t+1) - \mathbf{x}(t-m+1)]^t \mathbf{f}_k(t). \quad (80)$$

Combined with (75), this is enough to establish (79).

To prove (80), we proceed as follows: Note that

$$\begin{aligned} \Delta G_k(t) &= \sum_{\theta \in S(k, m)} E_k[\mathbf{x}(t - \theta_1 + 1), \dots, \mathbf{x}(t - \theta_k + 1)] \\ &\quad - \sum_{\pi \in S(k, m)} E_k[\mathbf{x}(t - \pi_1), \dots, \mathbf{x}(t - \pi_k)]. \end{aligned} \quad (81)$$

If  $\theta_1 \neq 0$ , then there exists a  $\pi$  in  $S(k, m)$  such that

$$(\theta_1 - 1, \dots, \theta_k - 1) = (\pi_1, \dots, \pi_k), \quad (82)$$

so that corresponding terms in the two summations cancel out. Similarly, if  $\pi_k \neq m$ , then there exists a  $\theta$  in  $S(k, m)$  such that (82) holds. So once again the corresponding terms in the two summations cancel out. Thus we are left with

$$\begin{aligned} \Delta G_k(t) &= \sum_{\theta_1=0} E_k[\mathbf{x}(t-\theta_1+1), \mathbf{x}(t-\theta_2+1), \dots, \mathbf{x}(t-\theta_k+1)] \\ &\quad - \sum_{\pi_k=m} E_k[\mathbf{x}(t-\pi_1), \dots, \mathbf{x}(t-\pi_{k-1}), \mathbf{x}(t-\pi_k)] \\ &= [\mathbf{x}(t+1)]' \sum_{(\theta_2-1, \dots, \theta_k-1) \in S(k-1, m-1)} \mathbf{e}_k[\mathbf{x}(t-\theta_2+1), \dots, \mathbf{x}(t-\theta_k+1)] \\ &\quad - [\mathbf{x}(t-m+1)]' \sum_{(\pi_1, \dots, \pi_{k-1}) \in S(k-1, m-1)} \mathbf{e}_k[\mathbf{x}(t-\pi_1), \dots, \mathbf{x}(t-\pi_{k-1})]. \end{aligned} \quad (83)$$

But both summations are the same, and equal  $\mathbf{f}_k(t)$  (see (58)). Hence we get

$$\Delta G_k(t) = [\mathbf{x}(t+1) - \mathbf{x}(t-m+1)]' \mathbf{f}_k(t), \quad (84)$$

which is precisely (80). This completes the proof.

#### 4. Conclusions

In this paper, we have formulated the problem of maximizing a general objective function over the hypercube  $\{-1, 1\}^n$  as that of maximizing a *multilinear polynomial* over  $\{-1, 1\}^n$ . Three modes of operation have been considered: asynchronous updating, gradient updating, and a new mode known as best neighbour updating. In the case of a quadratic objective function, the best neighbour updating does not appear to have been considered previously. We have shown that the asynchronous and the best neighbour updating rules converge to a local maximum of the objective function in a finite number of time steps. In the gradient mode, the behaviour of the network is not well-understood. For this purpose, we have *modified* the gradient updating rule in such a way that it incorporates both *temporal as well as spatial* correlations among the neurons. For the modified updating rule, we have shown that after a finite number of time steps, the network state vector goes into a limit cycle of length  $m$ , where  $m$  is the degree of the objective function. This does not preclude the possibility that the trajectory converges to a fixed point (which is a degenerate form of a limit cycle). If so, any such fixed point is a local maximum of the objective function.

While the modified updating formula presented here has the advantage that its dynamics are well-understood under gradient updating, it is natural to ask whether this advantage is enough to offset the added complexity of computing and implementing the modified updating rule. There can be no clear-cut answer to this question, as it is a matter of one's taste.

#### References

Hopfield J J 1982 Neural networks and physical systems with emergent collective computational capabilities. *Proc. Natl. Acad. Sci. USA* 79: 2554-2558

- Goles E, Fogelman E, Pellegrin 1985 Decreasing energy functions as a tool for studying threshold networks. *Discuss. Appl. Math.* 12: 261-277
- Bruch J, Goodman J W 1988 A generalized convergence theorem for neural networks. *IEEE Trans. Info. Theor* 34: 1089-1092
- Kamp Y, Hasler M 1990 *Recursive neural networks for associative memory* (Chichester: John Wiley)
- Hopfield J J, Tank D W 1985 'Neural' computation of decision optimization problems. *Biol. Cybern.* 52: 141-152
- Masti C L, Vidyasagar M 1991 A stochastic high-order connectionist network for solving inferencing problems. *Proc. Int. Joint. Conf. Neural Networks* Singapore, pp. 911-916
- Bruck J, Blum M 1989 'Neural networks, error-correcting codes, and polynomials over the binary  $n$ -cube. *IEEE Trans. Info. Theor.* 35: 976-987
- van Laarhoven P, Aarts E 1987 *Simulated annealing: Theory and applications* (Dordrecht: Reidel)
- Vidyasagar M 1985 *Control system synthesis: A factorization approach* (Cambridge, MA: MIT Press)



## Script recognition

P V S RAO

Computer Systems and Communications Group, Tata Institute of Fundamental Research, Bombay 400 005, India

MS received 11 May 1993; revised 25 February 1994

E-mail: rao@tifrvax.tifr.res.in

**Abstract.** This paper describes an approach for word-based on-line and off-line recognition of handwritten cursive script composed of English lower-case letters. The system uses simple and easily extractable features such as the direction of movement and curvature and the relative locations of regions where these suffer discontinuities.

Our approach was evolved based on our concept of 'shape vectors' introduced earlier. We visualise script characters as having shapes which are composed of comparatively straight segments alternating with regions of relatively high curvature. We derive the shape vectors from each script character essentially by identifying regions of least curvature and approximating these by straight lines. That these shape vectors carry adequate information about the identity of the character is established by showing that the original character can be faithfully reconstructed from the shape vectors.

We thus use slopes of the shape vectors and relative locations of points of maximum curvature (both highly quantised) as parameters for recognition. The system extracts parameters for individual characters from single specimens written in isolation and uses these to construct feature matrices for words in the vocabulary. These are used for matching with the feature matrices of test words during the recognition phase.

The advantage of the system is that it does not require elaborate training. Recognition scores are in the neighbourhood of 94% for vocabulary sizes of 200 words. The approach has been extended for off-line information as well and performs quite well even in this case.

**Keywords.** Character synthesis; cursive script; feature matrices; on-line and off-line recognition; overlapping segments; script recognition; shape vectors; tract segments; quantisation.

### 1. Introduction

The problem of script recognition by computers has been an active area of research over several decades. Cursive script recognition is, for obvious reasons, a much more complex problem than recognition of print or hand-printed characters.



Cursive script recognition involves deducing from the written script (which is a more or less continuous two-dimensional curve in an iconic form) the underlying sequence of discrete symbols or characters.

Cursive script information can be acquired in the on-line or off-line mode. In the former case, an electronic graphics tablet digitizes the pen coordinates continuously in real time, even as the characters are being written. Off-line systems use scanner digitizers to transfer a page image to the computer. Script recognition presents several difficult problems; we list the more important ones below.

(i) During cursive writing, strings of discrete characters are encoded into a more or less continuous curve. Decomposing this into character length segments is a non-trivial problem. Segmentation is eliminated if one does recognition at the word level directly, but then recognition will have to be accomplished at the word level; the number of classes increases very substantially. There is, in fact, a trade off between the complexities of segmentation and classification.

(ii) Proper choice of a small number of convenient parameters is an important step in script recognition. Individual script characters get considerably distorted in connected script, due to context effects caused by the continuous pen-down movement from one character to the next.

(iii) There are considerable differences in the handwritings of different writers. Even the same person writes individual characters and even words quite differently at different times. A recognition system, to be reasonably robust, has to be tolerant of such variations.

(iv) Word level recognition systems may use word level templates for direct comparison with the words in the test script. Alternatively, character templates can be used if it is possible to account for context effects; i.e. the distortions to the shapes of individual letters which occur in cursive writing.

## 2. Review of earlier work

Earlier script recognition systems have been based on statistical pattern recognition techniques. Mermelstein & Eden (1964a, 1964b) segment cursive script into a sequence of strokes. The strokes are characterised by velocity functions composed of pairs of quarter wave sinusoidal functions with different frequency and phase shift parameters and are classified into categories on the basis of topological similarities. There are rules constraining the adjacent occurrence of strokes and letters. Recognition accuracy is around 78%.

Riseman & Ehrich (1971) demonstrated the advantage of a structured approach: that contextual information helps to accomplish accurate word recognition even if character recognition performance is poor. Even with a character recogniser which divides characters only into five broad classes, word recognition scores could be as high as 98%, for a vocabulary of 300 seven-letter words. For longer words, performance would be even better.

Ehrich & Koehler (1975) adopted such a strategy for actual cursive script recognition. Here, the first down stroke of each character is recognised as a precursor to segmentation. Unlikely candidates are weeded out by using topological information. Substitution sets are generated at this stage. Context constraints are applied to narrow down the lists, based on two-class Neyman-Pearson decision

criteria. Error rates are low (less than 1%); however, the rejection rate for different writers could vary between 1% and 30% (for a dictionary size of 300 seven-letter words).

Farag (1979) used a stochastic model where each word is represented as a Markov chain of state transitions. He used eight types of basic strokes (straight line segments with slopes which are multiples of  $45^\circ$ ) of standard length. Words are segmented into sequences of such strokes. Learning consists in arriving at a set of stochastic forward transition matrices which define the transition probabilities between states for each word. Recognition uses a maximum likelihood classifier to determine which specific word might have produced the particular sequence of strokes. 100% recognition was achieved on a set of ten cursively written key words.

Sayre (1973) used topological features for word recognition. Rather than use stroke sequence information, he utilised letter bigram and trigram probabilities to eliminate improbable sequences. Word recognition scores were in the neighbourhood of 80%.

Bozinovic & Srihari (1989) adopted a multi-tiered approach for off-line recognition of cursive script. The problem of script recognition is dealt with in a series of transformations between different levels of representation. As a first step, the raw image is smoothed and slant correction is performed. The horizontal reference lines (which define the vertical extent of different types of characters) are then located. Minima in the 'y' component of the lower contours of the inscription are used as the basis for segmentation. The topology of the image is derived by a contour tracing operation. A description of the test pattern is then obtained, as the next step, in terms of certain relevant features and their locations. This yields a number of competing letter string hypotheses which account for the features in the test word pattern. The correct word is picked up on the basis of a best fit between the word string hypotheses and entries in the word lexicon. This system performed with accuracies ranging between 50% and 75% correct scores, for different writers, as the training data and the size of the lexicon are changed.

Kundu *et al* (1989) used a Hidden Markov Model (HMM) scheme for word recognition. They use a set of forty features which are nonredundant, easy to extract and independent of rotation, translation and size (e.g. number of loops, T- and X-joints and zero crossings, ratio of horizontal to vertical size, existence of isolated dots, and existence of semicircles as part of the character shape). Vector quantisation is used to arrive at an optimum set of vectors from a corpus of 2500 letters. Characters with similar shapes (such as 'e' and 'l') get grouped together in this approach. These classes are the states in the Hidden Markov Model at the word level. Statistical studies for the English language help to determine letter to letter and letter pair to letter transition probabilities. Each unknown letter in the test word is converted into a representative 'symbol' by means of the code book. The test word becomes a sequence of such symbols. The HMM framework helps to determine which lexical word is most likely to have yielded the observation sequence. The error rate is around 7.5%.

### 1.1 Comparison with other methods

In general, script recognition systems in the literature require elaborate training. This is due to the fact that such systems employ parametric representations which retain local shape information (along with deformations). In contrast to this, we implement an approach which does not deal with such deformations but with a representative (based on maxima and minima) pertaining to their normalised counterparts. These retain their global or archetypal characteristics, after all the deformations pertaining

to shape, orientation, positioning and size have been filtered out *a priori* (by heavy quantisation of slopes and coordinates). Since only canonical or archetypal shapes are used, training of the conventional type is dispensed with and robust recognition is possible.

### 3. Our approach

All the approaches described above use statistical methods of pattern recognition. They all need extensive training. Our approach, on the other hand, is knowledge-based, it seeks to exploit knowledge regarding the task for essentially dispensing with any extensive training.

#### 3.1 *Synthesis of cursive script characters*

In our earlier papers (Ramasubramanian & Rao 1988; Rao & Ramasubramanian 1991) we addressed the problem of synthesizing connected handwritten script from individual characters written in isolation (using the weighted average and the Bezier splicing techniques).

Our approach is to treat connected writing as a process of writing individual characters continuously, in the proper sequence, with minimal effort. In cursive script, the transitional link line between adjacent characters takes the form of a gradual anticipatory movement into the next character while the earlier one is still being written. The shapes of the individual character would get altered to a certain extent in this process. However, short of grossly sacrificing legibility, some deterioration in shape is tolerated in favour of smoothness and continuity of shape and movement. We synthesize these transition regions by the concatenation of individual character shapes to generate connected script. We make sure that continuity of motion and shape are preserved in the transition. This results in economy of movement; it also ensures in a smooth and efficient (i.e. minimal effort, minimal time) pen-down motion, as in the case of cursive writing by humans.

To facilitate synthesis, we divide each character into three segments: a prefix, a core and a suffix. The centrally located core (or shape identification) segments of adjacent characters are linked to each other by transition segments which are influenced by the suffix of the earlier character as well as by the prefix of the later one. Our synthesis consists essentially in the generation of the transition segments. The transition segments move gradually from the prefix of one to the suffix of another. We tried out two alternative, essentially equivalent, approaches: a weighted average method and a shape-specific Bezier splicing technique. Both were successful in generating cursive script. We could even replicate the distortion that occurs during rapid natural writing.

We demonstrated (Rao 1993, pp. 1–15) that even individual characters can themselves be visualised as being composed of (or realised as combinations of) simpler (straight line vector) elements, in the same manner as cursive script can be visualised as being composed of individual characters.

#### 3.2 *Decomposition of script characters*

We showed that we could segment characters using either of two criteria equivalently: minima in either the speed of movement or in the radius of curvature of the character

shape. Based on this, we made a conjecture that script characters can be visualised as resulting from an effort to trace in rapid succession a sequence of straight strokes or vectors. We called these strokes 'shape vectors' for the character.

We used a simple geometric construction procedure for fixing the slopes and lengths of the shape vectors and for identifying the suffix and prefix segments therein.

The technique that we used to generate cursive script from individual characters was successful in generating individual characters from the shape vectors; each loop or curved segment is generated by the concatenation of two shape vectors.

We thus demonstrated that shape vectors adequately characterise the canonical shape and identity of the original character. We therefore argue that they should provide a basis for script character recognition. Here, we describe an approach for recognition of connected script using parameters which relate to the shape vectors.

### 3.3 The recognition procedure

For cursive script recognition, we use features concerned with points of minimum radius of curvature (maximum writing speed) and maximum radius of curvature (minimum writing speed). In figure 1, the minima bear even numbers and the maxima are the points with odd numbers.

The parameters we use for representation are the  $X$  and  $Y$  coordinates and the radius of curvature. We quantise  $Y$  to four possible values:  $b$  and  $m$  (base line and midline) for characters such as  $c$  and  $n$  and  $h$  and  $l$  (high line and low line) for characters with ascenders and descender such as  $h$  and  $g$ . We measure  $X$  as the distance from the nearest minimum to the left. Figure 1 and table 1 illustrate the scheme.  $X$  is  $n$  for minimum 6 because it is very near to minimum 2 at its left, and  $f$  for minimum 10 since it is far from its left neighbour. The  $X$  value is useful in distinguishing between  $a$  and  $u$ ,  $g$  and  $y$ ,  $o$  and  $v$  etc. We define  $X$  as an increment rather than by its absolute value because the coordinates remain unchanged even after concatenating individual letters. We quantise curvature to two values:  $a$  for anti-clockwise and  $c$  for clockwise.

Maxima are represented by the slope angle, quantised as indicated below:

- (1)  $0 + 20^\circ$  : Right (R)
- (2)  $180 + 20^\circ$  : Left (L)
- (3)  $90 + 70^\circ$  : Up (U)
- (4)  $270 + 70^\circ$  : Down (D)

Fluctuations in slope due to varying slants in the angle of writing are eliminated by such coarse quantisation. Most maxima in script have values  $U$  or  $D$ . The last strokes of the letters  $v$ ,  $w$  and  $b$  have value  $R$ .

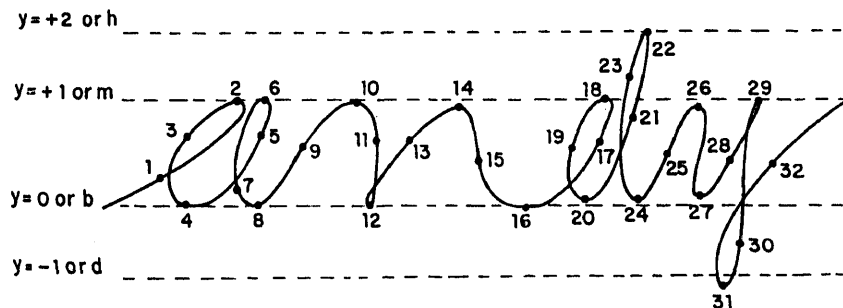


Figure 1. Maxima and minima for the script word 'andy'.

Table 1. Feature vectors for 'a' and 'n' of figure 1.

character a						character n					
Maxima	Serial No.	1	3	5	7	9	9	11	13	15	17
	Slope	U	D	U	D	U	U	D	U	D	U
Minima	X	f	f	n	f		f	f	f	f	
	Y	m	b	m	b		m	b	m	b	
	Curvature	a	a	a	a		c	c	c	c	
	Serial No.	2	4	6	8		10	12	14	16	

We extract parameters from single samples of individual letters. These constitute the feature vectors for the character. The feature vectors of characters a and n in figure 1 are shown in table 1.

As shown above, all the parameters are subject to heavy quantisation on a coarse grid. This quantisation filters out the fluctuations and shape deformations that occur in normal writing; only the maxima and minima are used to capture shape information. That our representation is adequate is clear from the high quality of resynthesis from the parametric representation. Variations in character sizes are taken care of by x and y quantisation while variations in slant angle are eliminated by slope quantisation.

We simply concatenate the feature vectors of individual letters to obtain the feature matrix of the word. In table 1, feature vectors for a and n are concatenated by merging the last column for a and the first column for n (both numbered 9).

We compile a lexicon consisting of all the words in the vocabulary from the feature vectors of individual characters. A similar feature matrix is obtained for the test word written by the subject, by analysing it in a similar manner. Recognition is accomplished on a best match basis, i.e. either the exact match or, failing this, the nearest match.

We use parameters relating to minima in our feature matrix. It might appear that this represents additional information not provided by the shape vectors. This, however, is not the case. The location, direction and magnitude of the shape vectors and the location of the junction point between the prefix and suffix segments are adequate to compute the location of the minima. Conversely, the lengths of the shape vectors are determined by means of a geometric construction that makes use of the location of the concerned minimum (Rao 1993, pp. 1-15). We use information regarding the minima (rather than equivalent information regarding shape vector size) purely for convenience; this is not a departure from our earlier stated approach.

#### 4. Details of implementation

##### 4.1 Data acquisition and preliminary processing

We acquire reference data and test data using a graphics tablet connected to a PC/XT compatible computer system. Reference data (not training data, in the commonly

understood sense of the word) is acquired for all 26 English lower-case letters written by the subject. This is used to compute feature matrices of individual characters. Test data (pertaining to the word to be recognized) are similarly acquired.

To facilitate extraction of height information, it is necessary to ensure that the script is in a straight line and that the characters are of a reasonably standard size. To this end, the upper and lower reference lines are provided on the writing surface as in ruled copy books for children. The subject is required merely to write broadly within these lines. There is no need for anything like copy book precision.

As mentioned already, we can locate maxima and minima in the character curve on the basis of the radius of curvature or the speed of movement of the pen. Once these are located, the relevant parameters are extracted and quantised easily. We used three different methods (all dependent on the determination of the radius of curvature) for locating these.

Firstly, we can use the analytical expression for radius of curvature of the character in terms of the 'x' and 'y' components of the pen velocity and acceleration, using the expression,

$$(x'y'' - y'x'')/[x'^2 + y'^2]^{3/2}.$$

Secondly, the curvature ' $\rho$ ' is also the arc derivative of the slope angle, ' $\theta$ '. We can therefore compute ' $\rho$ ' as a mean over ' $2n$ ' points using the relationship

$$\rho(j) = [\theta(j+n) - \theta(j-n)]/(2n.t.v).$$

Here, ' $t$ ' is the sampling period, ' $v$ ' is the pen speed and ' $\theta(n)$ ' the slope angle of the curve at the  $n$ th sampling point.

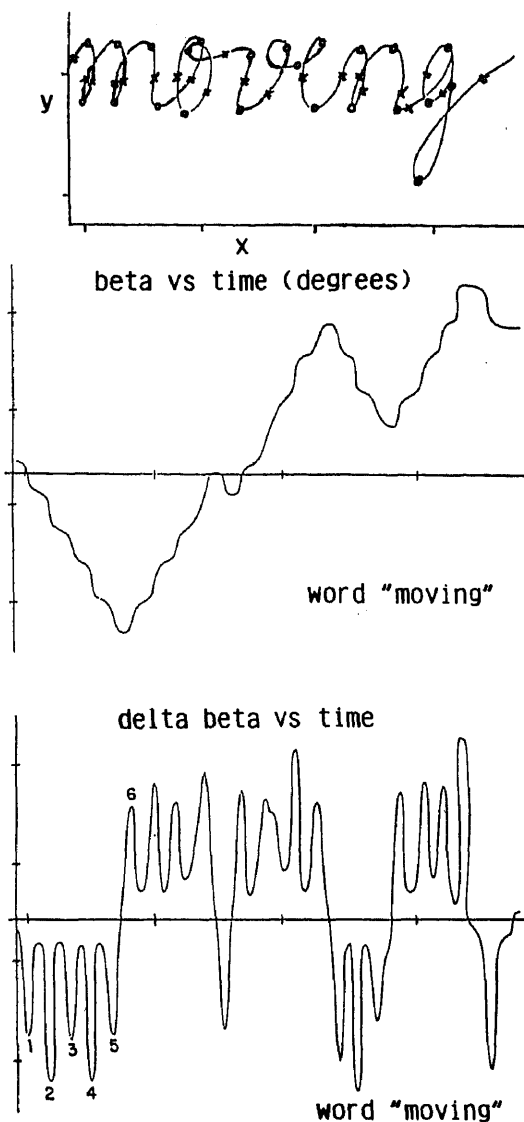
For computational simplicity, ' $v$ ' can be taken to be constant and be dropped from the denominator; ' $n$ ' and ' $t$ ' being constants, the curvature can be taken as being proportional to

$$[\theta(j+n) - \theta(j-n)].$$

Pen speed ' $v$ ' of course, is minimum for maximum ' $\rho$ ' and maximum for minimum ' $\rho$ '. If we assume this to be constant, our apparent minima will be higher and maxima lower. Despite this, we are able to locate the maxima and minima very effectively (see figure 2). The maxima in pen speed are minima in radius (i.e. our minima).

Thirdly, it is easy to locate minima (and maxima) using the fact that the stylus travels fast and almost in a straight line near the maxima but is slow and almost doubles back on itself as the curve turns nearly  $180^\circ$  around that minima. The chord distance (not the arc distance) between two sample points separated by constant time would therefore be minimum around minima and maximum around maxima.

Handwritten characters are usually not vertical in their orientation; they tilt to a greater or lesser extent, usually to the right. Even a left tilt, though less frequent, is not uncommon. This tilt or slant remains more or less invariant for each person and is one of the features that characterise an individual's handwriting. It would seem very desirable to eliminate this slant so that we can deal with more standard - essentially vertical - characters. We did this by means of a number of slant estimation and correction processes but found that, except for extreme slants, this is not really essential, since our method of parameter representation and quantisation for the script characters is robust enough to tolerate even major slants in the writing.



**Figure 2.** Radius of curvature against time for the word "moving" by the slope slant method.

#### 4.2 Word dictionary compilation

We assemble each word in the dictionary by linking parameters of the constituent characters in a sequence. It is important to avoid duplicate or redundant maxima in this process. Each character has two maxima, one at each extremity. During concatenation, of two characters, the last maximum of the previous character and the first maximum of the next character are merged and one maximum (the second) is eliminated.

Some subjects tend to start with a minimum on the base line while writing isolated words or characters. We discard such spurious minima during concatenation. We also provide for certain letters being written in two or more different ways; we treat

the different variants of a character as different characters. The dictionary also includes multiple versions of words using such letters.

#### 4.3 *The recognition procedure*

It is important that the procedure for comparing the extracted features of the input word with the pre-stored features of the dictionary words be efficient. We narrow down the list of prospective candidates as far as possible on the basis of a rough match and perform detailed matching only within this short list. We use the number of minima in a word as the criterion for short listing. We reject all dictionary words where the mismatch in the number of minima exceeds three. A detailed matching of individual words is attempted only for words where the mismatch is smaller.

If an exact match is found within the short list, recognition is successful. Failing this, we select the best match, using the following procedure. We impose a penalty score for each non-matching minimum ( $-10$  points) and each non-matching maximum ( $-2$  points). We choose the word in the short-list with the minimum penalty score. Recognition fails if either the short list is empty or if, for all words in the short list, the mismatch during detailed matching turns out to be excessive.

We use a dynamic programming method of comparison to allow for the insertion of spurious maxima or deletion of genuine ones.

#### 4.4 *Ambiguities near the extremities*

A few extra points might be digitised when the stylus moves away from the paper after writing a word. This might look like a small curve in an entirely different direction. In normal script, on the other hand, there are no discontinuities close to the end of a character. We eliminate such discontinuities (and any points that follow such a discontinuity). We discard the first minimum of the test word if it lies on the base line (since no character and hence no word has its first minimum on this line). Also, we ignore the last maximum of the word, since this may not always be present in normal writing.

We use finite difference methods throughout to ensure immunity to noise. This means however that we might miss features very near the start and the end points, if the subject uses very short prefixes and suffixes while writing isolated characters during the reference data acquisition phase. This is true even in the case of start and end segments for test words. Also, in view of the quantisation procedure adopted for the positions of the minima and maxima, there could be problems if the subject writes the characters in a word too close to one another. We did not encounter such problems in our experiments.

#### 4.5 *Missed and spurious minima*

Occasionally, we may miss a minimum: e.g. if the lower loop of the letter 'y' or 'j' is almost circular. We introduce a minimum between two existing maxima, if the slope changes over a range of say  $220^\circ$  without the occurrence of a minimum.

Spurious minima may occur due to wiggles or tremors in writing and we use a few simple techniques to eliminate them. Firstly, we fix thresholds and ignore all maxima and minima whose values are below these. Secondly, we expect minima to be physically well separated along the length of the curve. Hence, if two consecutive



minima occur too close to each other (separated by less than a fixed number 'n' of sample points), then we discard the second. Alternatively, we can measure the difference in slope angles on either side and check this against a specific (minimum) threshold. Also, there are specific locations where only minima can occur. We define such regions in relation to the reference lines. We discard minima which do not satisfy these criteria as being spurious. We found each of these techniques to be quite effective in eliminating spurious minima.

#### 4.6 Ambiguous segments

Points of inflection, i.e., points where the curvature crosses zero, are often misleading. For example, we expect the downstroke for 'f' to be anticlockwise throughout. However, there is in some cases a momentary shift to clockwise movement midway and then back again. These, however, get eliminated by the criteria we use.

We give low weightage to minima and the corresponding maxima existing on the transition segments in the dictionary word, especially those pertaining to characters with a horizontal suffix such as 'b', 'o', 'v' and 'w'. This is because, in general, this transition shows wide variability.

### 5. System performance

We established the feasibility of the overall approach in a preliminary experiment which used a minimal set of parameters: the sign of the curvature at the minima and the direction of movement at the maxima (up, down, right and left). We could accomplish word recognition even with such drastic data abstraction. Understandably, there is confusion between similar letters such as 'e' and 'l', 'u' and 'a' etc.

We used the following in a second experiment:

- (i) *for the maxima*: location (normalised and quantised) and direction of movement (clockwise or counter clockwise);
- (ii) *for the minima*: direction of movement (quantised: up, down, left and right).

The results we obtained with this set of parameters are summarised in table 2.

### 6. Script recognition in the off-line mode

#### 6.1 Off-line to on-line conversion

We assume that off-line information regarding the character of word to be processed is collected and made available in the raster scan mode. Our object is to organise

**Table 2.** Recognition scores for experiment 2.

	Run 1	Run 2	Run 3
Vocabulary size	24	67	67
Correct words recognised by exact match	20	39	57
Correct words recognised by nearest match	04	24	09
Number of words incorrectly recognised	nil	04	01
Recognition score	100%	94%	98%

the 'points' so provided in each scan line into individual tracks or sequences of adjacent or connected points. We then merge these into a single track i.e. a sequence of points which recreates the time order in which the curve has been traversed during writing.

6.1a *Formation of track segments:* At each point, a decision has to be made whether it is part of an earlier track. The criteria for this are:

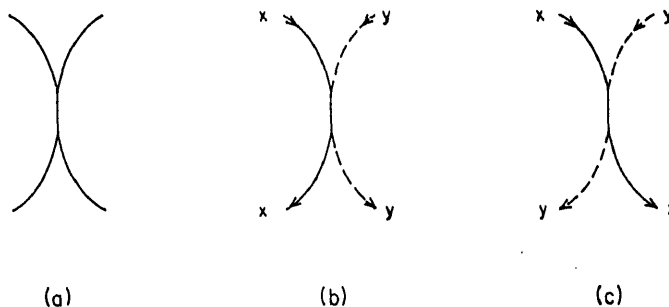
- (i) a distance threshold – that it be within a certain distance of one of the two ends of the track.
- (ii) a slope threshold – that it be within a certain distance from the expected position of the point, as computed by linearly extrapolating the track.

We use only the distance threshold in case of single point tracks. We carry out extrapolation using the slope criterion based on the average slope computed using the previous three or four points in the track. We take a new point to be a part of an earlier track if these criteria are satisfied, and add it to that track; else, we form a new track.

This procedure yields track segments composed of between 10 to 20 points each. A few very small segments (of less than three points) also remain. We absorb each such segment into a large track if this can be done by inserting a point in the gap, but only provided that after this insertion, distance and slope matches are within specified limits.

6.1b *Formation of tracks:* We need to combine the track segments so formed into one track. This is done as follows. We compute the degree of mismatch between two tracks pair-wise, as a linear function of the position mismatch and slope mismatch at the joining point. We also join track segments, pair-wise, best match first; if the procedure is successful, we will eventually end up with a single track.

This method works quite well even when the track crosses itself. There is a problem, however, when two segments of the track overlap, merging into one. Track consolidation cannot be completed in such a case because, two segments having merged, we would be one segment short. Also, there would be two different ways of connecting the segments (see figure 3); we resolve this ambiguity by using the property that points of inflection are quite rare in cursive script; and hence that track curvature



**Figure 3.** (a) Overlapping segment. (b) & (c) Two different ways of joining the segments. Case (b) (where the curves remain clockwise on either side of the common segment) is chosen; (c) is very unlikely.

**Table 3.** Recognition with off-line and on-line data: total number of words 47.

	Correctly recognised		Wrongly recognised	Connection failure
	Exact match	Nearest fit		
Raw on-line data	39	7	1	Not applicable
Reconverted data (on-line to off-line to on-line)	20	15	5	7

does not change sign (from clockwise to anti-clockwise or vice versa) on the two sides of the common segment. Thus, we link the two clockwise segments together, as also the two anti-clockwise segments.

### 6.2 Comparison of off-line and on-line performance

We use the following procedure to evaluate the performance of the off-line to on-line conversion procedure. We collect word data on-line and convert it into off-line raster scan mode by a fairly simple procedure. This procedure introduces errors due to quantisation. We provide the raster scan information so generated as input to the off-line to on-line conversion procedure.

This original and reconverted on-line data sets are independently used for recognition. Table 2 shows the results.

These results demonstrate the effectiveness of the off-line to on-line conversion procedure. It may be noted that while off-line recognition scores do drop, this is due more to connection failures rather than due to misrecognition. In other words, the system tends to err on the safer side of not connecting tracks rather than connecting them wrongly.

### 6.3 Problems in off-line to on-line conversion

An analysis of the failures of track connection reveals the following problems.

- (i) Due to slow pen movement, the density of points is very high near the beginning and end of each word. This makes track creation difficult. A solution is to eliminate short segments at the beginning and end of each word. This seems to work in most cases. This problem, it may be noticed, would exist only when off-line data is created from data acquired on-line and will not occur when information is acquired directly from a scanner.
- (ii) Where tracks turn sharply, track segments might be assembled in the wrong order.
- (iii) We have already discussed the problem of overlapping segments.

## 7. Discussion and conclusions

### 7.1 Knowledge based system

The statement that our system is knowledge-based needs justification. This knowledge in our system is implicitly contained in the model we propose for synthesis of cursive

script (from individual characters) and of the characters themselves (from shape vectors) and in the parameters chosen for representing the signal.

Our system is able to operate without training because it incorporates task domain knowledge in its structure. In a manner of speaking, all recognition systems require knowledge regarding the task domain in order to perform recognition successfully. Statistical systems start without any a priori knowledge. Such classifiers acquire this knowledge in a process of elaborate training. Hidden Markov Model and neural nets incorporate some knowledge about the task domain in terms of the exact topology chosen for them. Since, however, this knowledge is minimal, such systems still need extensive training. A knowledge-based system, on the other hand, incorporates a significant amount of task specific knowledge and, as a consequence, is able to operate with minimal training (or even without any training at all). Our system is knowledge based in this sense.

Task domain knowledge is known to be useful for signal estimation in the presence of noise: e.g. cleaning a scanned picture of a script word into an  $x - y$  displacement representation of pen tip movement, time domain representation of the speech signal into spectrographic, formant trajectory or vocal tract area function type of representations. The following paragraph illustrates how this knowledge is useful even for our own recognition task.

Script recognition in real life gets complicated due to context effects, sloppy writing, inter-subject variability, etc. Most recognition systems (e.g. HMM systems; Kohonen nets) seek to model these distortions, delineate class distributions (e.g. by modelling the probability densities) and select class prototypes. We take into account the fact that there exists a notional ideal pattern (for each script character or word), which the writer is attempting to generate while he writes. The distortions that occur in actual writing can then be treated as added noise. Our aim, in a sense, is to map the noise-added version into the copy book counter part and this is accomplished in our system. We do this in two stages. To start with, we use a parametric representation that ignores the actual and detailed shapes of the character. We, thus, are able to filter out the deformations and retain only the broad shape features. In addition, we quantise the values of the parameters in a coarse grid, which filters out fluctuations due to character positioning, slant orientation, size and spacing.

On the other hand, conventional pattern classifiers would have operated on the parameter matrix and estimated the mean, variance and spread by means of an extended training phase, accomplishing recognition by a distance criterion.

A major advantage of our recognition approach is that it does not require a training phase. The dictionary words need not even be fed in the script form. We need merely to 'initialise' the system by feeding it one sample each of the 26 lower case script letters of the alphabet. The robustness of the method is amply demonstrated by the fact that performance does not degrade significantly even when the initialisation (training) and test samples are written by different subjects.

In its present form, this method is very suitable for vocabularies in the range of upto a few hundreds of words. Currently we make a straight comparison between the feature vectors of the test word and the dictionary words; we thus pay a heavy penalty for spurious (or missed) maxima and minima because these will cause the remaining part of the word to be misaligned. With more elaborate dictionary match methods (e.g. stack-search algorithms), the system can be used even with significantly larger vocabularies. The features extracted by the present system can be used for neural network and HMM techniques. Preliminary investigations in this direction are

very promising. The recognition scheme works very well for on-line inputs. It imposes only a few acceptable constraints.

While the principles underlying the present approach are general enough to be valid for all cursive scripts, they are particularly well suited for the Roman scripts. With some changes, it should be possible to apply them for machine recognition of cursive scripts in, say, Indian languages.

The scheme has been tried so far mainly with data acquired in the on-line mode. It has been shown that it is capable of being extended even for the recognition of page scan data, by converting this information into the on-line mode. The results obtained demonstrate the feasibility of the approach, and the potential for improvement in performance with further finetuning of the techniques.

## References

- Bozinovic R M, Srihari S N 1989 Off-line cursive script word recognition. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-2: 68-83
- Ehrich W, Koehler K J 1975 Experiments in the contextual recognition of cursive script. *IEEE Trans. Comput.* EC-24: 182-195
- Farag R F H 1979 Word-level recognition of cursive script. *IEEE Trans. Comput.* EC-28: 172-175
- Kundu A, He Y, Bahl P 1989 Recognition of hand-written word: first and second order hidden Markov model based approach. *Pattern Recogn.* 22: 283-297
- Mantas J 1986 An overview of character recognition methodologies. *Pattern Recogn.* 19: 425-430
- Mermelstein P, Eden M 1964a A system for automatic recognition of handwritten words. *Fall Joint Comput. Conf., AFIPS Conf. Proc.* 25
- Mermelstein P, Eden M 1964b Experiments on computer recognition of connected handwritten words. *Inf. Control.* 7: 255-270
- Ramasubramanian V, Rao P V S 1988 Connected script synthesis by character concatenation - An overlap and weighted average formulation. *Comput. Sci. Inf.* 19: 1-10
- Rao P V S 1993 Shape vectors: an efficient parametric representation for the synthesis and recognition of hand script characters. *Sādhana* 18: 1-15
- Rao P V S, Ramasubramanian V 1991 Connected script synthesis by character concatenation - A Bezier curve based formation. *Inst. Electron. Telecommun. Eng.* 37: 485-493
- Riseman E M, Ehrich R W 1971 Contextual word recognition using binary diagrams. *IEEE Trans. Comput.* C-20: 397-403
- Sayre K M 1973 Machine recognition of handwritten words: A project report. *Pattern Recogn.* 5: 213-228

## Solid finite elements through three decades

D N VENKATESH and U SHRINIVASA\*

Department of Mechanical Engineering, Indian Institute of Science,  
Bangalore 560 012, India

E-mail: [udipi@mecheng.iisc.ernet.in](mailto:udipi@mecheng.iisc.ernet.in)

**Abstract.** Conventionally, solid finite elements have been looked upon as just generalizations of two-dimensional finite elements. In this article we trace their development starting from the days of their inception. Keeping in tune with our perceptions on developing finite elements, without taking recourse to any extra variational techniques, we discuss a few of the techniques which have been applied to solid finite elements. Finally we critically examine our own work on formulating solid finite elements based on the solutions to the Navier equations.

**Keywords.** Solid finite elements; extra-variational techniques; interpolation functions; convergence criteria; Papcovich–Neuber solution; Navier equation.

### 1. Introduction

Development of the finite element method (FEM) as an analysis tool for continuum problems coincided with the advent of powerful digital computers. Using this method it is possible to establish and solve equations pertaining to complex systems in a very simple manner. It is precisely because of these reasons that today FEM has come to stay as a powerful tool in engineering analysis and design encompassing many diverse fields including structural mechanics, fluid mechanics, solid mechanics, electromagnetism etc. In fact it has emerged as a very popular analysis tool in interdisciplinary problems. The popularity of the method can also be attributed to the ease with which complex domains can be handled, requiring no additional techniques.

As in the case of all original development it is difficult to pinpoint exactly when FEM was discovered. It could be attributed to three separate groups – mathematicians (Courant 1943; Collatz 1950; Courant & Hilbert 1953), physicists (Synge 1957) and engineers (Turner *et al* 1956). Very good surveys of the origins of FEM exist and are today commonly found in introductory chapters in many textbooks (Zienkiewicz 1973; Yang 1986). As is often the case with diverse groups working far removed from one another, emphasis on the various aspects of the study has been different for both engineers and mathematicians. Engineers were not deterred by the lack of an elegant mathematical theory, while mathematicians concerned themselves with those aspects of the problem which interested them more, like convergence.

Today, FEM has been put on a relatively more sound mathematical footing. Extensive literature exists on its mathematical foundations and many aspects can be found elegantly represented in books (Ciarlet 1978; Oden 1983). FEM (as used by both mathematicians and engineers) involves a series of approximations. While it has generally been accepted that any procedure of discretization will involve a series of approximations, the approximations of FEM are more far reaching than others. For example, the breaking up of an elastic continuum itself is an approximation, for these elements are joined together only at the nodes.

Amongst the many other approximations made by the method, the most crucial one appears to be that of assuming that the displacements within the element can be expressed as a linear function of the nodal point displacements. This seems to have an overriding impact on the formulation of successful finite elements. We call this functional relationship from now on as the "interpolation function". This relationship is also referred to by various names in literature like shape functions, basis functions etc.

The choice of the geometry of the element and the interpolation function leaves a great deal of scope to the ingenuity and skill of the engineer designing finite elements. Solutions obtained would obviously depend a great deal on the exercise of this skill. FEM is normally implemented as the minimization of a functional. For example the popular displacement based approach, uses the principle of minimization of the total potential. The solutions obtained on minimization of the functional would then be obviously constrained by the choice of the assumed displacement field. Various types of interpolation functions are used in the literature. There have also been many methods and functionals used for formulation of finite elements. It is beyond the scope of this article to examine all of them. We shall therefore restrict ourselves to examining the methods and interpolation functions used in the formulation of three-dimensional finite elements for elastostatics.

## **2. 3-Dimensional element formulations and interpolation functions**

Conventionally finite elements are formulated by strictly adhering to three cardinal principles, known as the "convergence criteria".

### **2.1 The convergence criteria**

The assumed interpolation functions limit the infinite degrees of the system to a function with finite degrees of freedom. Therefore the minimum of the functional will not represent the correct equilibrium configuration as it would largely depend on number of degrees of freedom chosen. This will be true in spite of the fineness of the finite element mesh and the number of subdivisions used. Therefore, conventionally, in order to ensure convergence to the correct solution, certain criteria are required to be satisfied by the interpolation functions. They are the following (Zienkiewicz & Taylor 1989).

- (1) "The displacement function chosen should be such that it does not permit straining of an element to occur when the nodal displacements are caused by a rigid body displacement."
- (2) "The displacement function has to be of such a form that if nodal displacements

are compatible with a constant strain condition such constant strain will in fact be obtained."

- (3) "The displacement functions should be so chosen that the strains at the interface between elements are finite."

Of course all the three criteria are needed to be satisfied only in the limit when the element sizes shrink to zero. Earlier they (or slightly modified versions of them) went by the names commonly referred to as continuity, compatibility, and a third criterion of completeness. Mathematical statements encompassing them exist in the name of "functional completeness" (de Arrantes Oliveira 1968; Strang & Fix 1973). Unfortunately they do not speak about the rates of convergence, except recognizing that in the limit as element sizes go to zero (thereby calling for repeated mesh refinement), solutions converge. In fact it can be shown that all elements which have been formulated by rigorously adhering to the above criteria will converge (Strang & Fix 1973). However, these criteria could not explain the phenomenon of locking and the resulting loss in convergence.

## 2.2 Solid elements satisfying the convergence criteria

We first consider solid elements developed by adhering to the above convergence criteria.

**2.2a Tetrahedral elements:** The first formulation of a simple tetrahedral element was by Gallagher *et al* in 1962 and was used in the stress analysis of heated complex shapes. Early elaborations of tetrahedral elements were by Melosh (1963), Argyris (1965) and Clough (1969). One of the early extensive numerical studies is due to Rashid (1969, 1970).

In these studies (and future elements to be discussed) tetrahedral elements were formulated in terms of volume coordinates similar to the triangular elements which were formulated in area coordinates, and were simple generalizations of the triangular elements. These elements were numerically integrated to obtain the element matrices.

As can be easily deduced the first of the tetrahedral elements (Gallagher *et al* 1962) was a  $C^0$  continuous, 4-node, 12-degree-of-freedom (d.o.f.) constant strain tetrahedron, with linear shape functions along the three orthogonal cartesian directions.

Clough (1969) used a  $C^0$ , 10-node, 30 d.o.f. linear strain tetrahedron by adding midside nodes. This tetrahedron used complete quadratic polynomials in the three directions. Other higher order elements have also been formulated. Argyris *et al* (1968b) obtained a  $C^0$ , 20-node, 60 d.o.f. quadratic strain tetrahedron by adding nodes at the one- and two-third points of each of the six sides of the tetrahedron. The interpolation function was a complete twenty-term cubic polynomial in volume coordinates. Rashid *et al* (1969) used a  $C^0$ , 16-node, 48 d.o.f. tetrahedron by omitting the centroidal face nodes. Argyris *et al* (1968c) proposed the TEA8 element with 8 nodes and 60 d.o.f. This element has four centroidal nodes in addition to the four vertex nodes. Each vertex node has the following degrees of freedom:  $u$ ,  $\partial u/\partial x$ ,  $\partial u/\partial y$ ,  $\partial u/\partial z$  and similar d.o.f. in the  $v$  and  $w$  directions. Hughes & Allik (1969) and Fjeld (1969) have formulated and used a 4-node, 48 d.o.f. tetrahedron by using the four vertex nodes and with d.o.f. of  $u$ ,  $v$ ,  $w$  and their derivatives in the  $x$ ,  $y$  and  $z$  directions at each node. Being a higher order element with derivative degrees of freedom, this element required higher order continuity, and as could be expected "this is the most advantageous tetrahedron introduced" (Yang 1986) till its date of publication.



Comparative studies on the above tetrahedral elements abound (in a partial sense – upto their dates of publication) and the interested reader is referred to the work of Fjeld (1969). Initial work regarding the development of tetrahedral elements were hampered due to the manual input of data required. With the advent of high speed computer graphics many algorithms can be found (Pammer & Szabo 1981; Nguyen 1982) which aid the user in developing the data for the analysis.

**2.2b Rectangular hexahedral elements based on the convergence criteria:** Initial work on conforming hexahedral elements tended to be restricted to rectangular ones. This could be attributed to two reasons. (a) The generalized coordinate approach could yield geometries which were not invertible and (b) since the faces and sides of the rectangular elements are orthogonal to one another, such elements can be formulated using non-dimensional local coordinate systems. A number of such elements have been formulated. Amongst the first of these was a  $C^0$ , 8-node, 24 d.o.f. linear displacement rectangular hexahedron (Melosh 1963; Clough 1969). The element used "trilinear" displacement interpolation functions in the three orthogonal directions. The stiffness matrix was explicitly given by Melosh (1963). The addition of one node to the midpoints of each of the 12 sides of the hexahedron gives a  $C^0$ , 20-node, 60 d.o.f. quadratic displacement hexahedron. Just like the trilinear element used incomplete cubic polynomials, this element used incomplete quartic ones. This element was first enunciated by Clough (1969) and highlighted by Rigby & McNeice (1972) and Pawsey & Clough (1971). The addition of two nodes, at the  $1/3$  and  $2/3$  points of the edges yields a  $C^0$ , 32-node, 96 d.o.f. cubic displacement rectangular hexahedron, which uses incomplete quintic polynomials, see Ergatoudis *et al* (1968). Other higher order elements have also been formulated. The addition of four facial nodes on each of the 6 faces of the hexahedron in addition to the 32-node hexahedron above and the addition of 8 interior nodal points yields a 54-node, 192 d.o.f.  $C^0$  hexahedral element, first used by Argyris & Fried (1968). Here the interpolation function can be obtained by taking the product of 3 cubic polynomials in 3 directions. This of course leads to an incomplete higher-order field (such elements are known as Lagrangian elements). Another commonly used rectangular Lagrangian element is the one obtained by taking the product of quadratic polynomials in three orthogonal directions to obtain a 27-node, 54 d.o.f. hexahedron with a centroidal node. Since no loads could be applied on the centroidal node, the d.o.f. corresponding to this node are statically condensed.

The Lagrangian elements have a disadvantage in that the interpolation functions require the use of unusually large degrees of the polynomial for interpolation.

The use of first derivative degrees of freedom (along the 3 directions  $x$ ,  $y$  and  $z$ ) yields a 8-node brick with 12 d.o.f. at each node (96 d.o.f.). For such an element the displacement functions could be assumed in the form of Hermitian polynomials as in Argyris *et al* (1968b) or an incomplete fifth order field as used by Zienkiewicz *et al* (1970).

On account of the fact that 3-dimensional elements require large matrices to be stored and enormous computational effort, numerical experiments concerning these solid elements have by and large been "spotty" (Yang 1986) as compared to their 2-dimensional counterparts. But it is obvious from the above discussion that a large number of elements of both tetrahedral and hexahedral geometries can be formulated. Some numerical comparisons do exist in literature. The interested reader is referred to studies by Melosh (1963), Rashid (1970), Hughes & Allik (1969), Fjeld (1969),

Chacour (1972) and Ferguson & Clark (1979). For a brief idea (in tabular form) of the relative performance of each element (accuracy/unit computational time) the reader is referred to Yang (1986).

One other hexahedral element which has been formulated, though not frequently used is a fourteen-noded brick obtained by the use of 6 facial nodes in addition to the 8 vertex nodes. This element suffers from the drawback that it is very difficult to obtain a frame invariant element. Irons & Ahmad (1979) have noted that even though the element was very promising as it would produce a small bandwidth/frontwidth for analysis "there is something wrong with this element". Kidger (1990) and Smith & Kidger (1992) using averaging along three directions have produced 14-node hexahedral elements and obtained success with some problems.

**2.2c Other elements:** Finite elements of shapes other than tetrahedra and hexahedra have also been formulated. Some of these are wedge-shaped and pentahedral elements. For example, in the wedge-shaped (triangular prisms) the interpolation functions can be obtained as a product of the Lagrange approach and the serendipity approach. See Zienkiewicz (1977) for the interpolation functions used in formulating the 6-node-18 d.o.f., 15-node-45 d.o.f. and 26-node-78 d.o.f. triangular prism elements. Many forms of pentahedral elements also exist. For one of these elements the reader is referred to Haggemacher (1993).

### 2.3 Solid elements based on hierarchic interpolation functions

So far in all the elements mentioned above, we find that we need to increase the number of nodes when we want to increase the order of the interpolating polynomial or alternately use elements with derivative degrees of freedom. Yet another means of generating interpolating functions for elements is to use hierarchic approximations. In this method one needs to associate the monomial term in each interpolating polynomial with just a parameter and not to one with an obvious physical meaning. The only constraint associated with these hierarchic functions is that they need to have zero values at the end of the range – in this case the vertices or nodal points along the edge under consideration. Using these polynomials one can arrive at a variety of interpolation functions for elements of different geometries. In fact it is possible to obtain a general form for interpolation functions, as demonstrated by Peano (1976). For 3-dimensional elements "a simple identification of the hierarchic parameters on the interfaces will automatically ensure  $C^0$  continuity of the approximation" (Zienkiewicz & Taylor 1989). In order to obtain optimal forms of hierarchical functions – those that result in a diagonal equations system, it can be shown that such interpolations can be obtained using orthogonal polynomials, for example Legendre polynomials. The products of these interpolation functions in three directions can yield interpolation functions useful for 3-dimensional elements (Zienkiewicz *et al* 1983).

### 2.4 Isoparametric elements

It is obvious from the above discussion, that it is possible to generate a large number of solid finite elements very easily. Despite the ease of formulation, all the above mentioned elements suffer from one particularly severe drawback – their poor curve fitting ability. This ability is severely tested when modelling real life problems with complex curved geometries.

In order to overcome this drawback isoparametric elements are used. The method allows for the use of a nonlinear transform, mapping of a given finite element geometry to a "parent" one. "The interpolation of the element coordinates and the element displacements using the same interpolation functions, which are defined in the natural coordinate system, is the basis of the isoparametric element formulation" (Bathe & Wilson 1976). In addition another particular drawback of the generalized coordinate method is also overcome, namely, the excessive care which is to be exercised in order to express the generalized coordinates in terms of nodal point displacements involving the inversion of a matrix (which for certain geometries could be singular). Also it overcomes the increased computational effort required in the generalized coordinate approach to transform element matrices from the element (local) coordinate system to a global one – this is more easily done in the case of isoparametric elements. These advantages are achieved by the direct use of interpolation functions used for the displacement. Isoparametric elements are today very popular. The first isoparametric element was developed by Taig (Robinson 1985) in 1958, but the first work was published by Irons (1966). Since then many publications comparing one set of isoparametric elements with another and highlighting the advantages of these with respect to the generalized coordinate elements have appeared (Ergatoudis *et al* 1968; Clough 1969; Zienkiewicz *et al* 1969; Ahmad *et al* 1970). The interpolation functions used in the formulation were similar to those described in the previous section, except that interpolation functions for the displacement field were directly used.

The basic problems of using isoparametric elements (as in non-isoparametric formulations) are two fold (Zienkiewicz *et al* 1971; Pawsey & Clough 1971; Wilson *et al* 1973). "Firstly excessive shear strain energy is stored in these elements and, secondly, as the element becomes "thin" (for example, very large  $l/d$  ratios in the case of analysis of plates and shells) the stiffness coefficients corresponding to the transverse degrees of freedom are larger than those of longitudinal displacements, which results in numerical ill-conditioning" (Bathe & Wilson 1976).

### 3. Extra-variational techniques

Most of the elements mentioned in the above sections were derived by adhering strictly to the convergence criteria (except the ones using hierarchic interpolations). The behaviour of these elements in situations – such as bending or near incompressibility (especially the lower-order elements) left a lot to be desired. These problems, depending on the type being solved and the element under consideration, are known by various names. Three-dimensional elements are known to suffer from delayed compressibility locking respectively. Here, the term "locking" will be used to "denote an indefinite decay of accuracy in displacement recovery" (Naganarayana 1991). Locking is prevalent in other structural problems also, like shear locking in flat plates/shells (see Bathe & Dvorkin (1985) and Donea & Lamain (1987)) and membrane locking in curved beam and shell structures (Stolarski & Belytschko 1981; Prathap 1985). Other common problems encountered are "violent stress oscillations" (Prathap 1992) and "delayed convergence" (Naganarayana 1991). It will not be wrong to suggest that the various new elements which are formulated lately address themselves to tackling these problems. In fact there does not appear to be any solid element which

is totally divorced from these problems. Since the early days of the finite element method, the development of such elements has been the source of both a challenge as well as a motivation for element developers.

Many techniques do exist in literature to get around these problems rather than attempt to solve them. (Consistency (Prathap 1992) is an attempt at solving the problem, though not fully successful for solid elements.) These techniques have been labelled variously as "tricks" (Prathap 1992) by some or "variational crimes" (Strang 1972) by some others, for they do not strictly adhere to the exact rules laid down by the variational principles on which the mathematical model is based. Many of these techniques can be categorized as "*ad-hoc*" (Naganarayana 1991), for their success in some problems does not necessarily imply the same when extrapolated to other problems. The list of these techniques is very large and continues to grow even today. "The task of developing good finite elements never seems to be finished. Designers return, again and again, to the same basic configuration of nodes and find some way to eke out an improvement" (MacNeal 1992).

It would be a Herculean task to list the entire set of these techniques individually. We list below only some of the above mentioned techniques, representing a good sample of select procedures which over the years have come to stay – these techniques have been regarded as the "milestones" (MacNeal 1992) of progress in the FEM.

(1) Reduced/selective integration: This is among the first of the so called "variational crimes" discovered. In this method the strain energy is not integrated exactly. For example, the 8-node solid element is integrated using a  $2 \times 2 \times 2$  Gauss rule and the 20 node brick by a  $3 \times 3 \times 3$  Gauss rule. The 27 node Lagrangian element uses a  $3 \times 3 \times 3$  order rule. It is well-known that an  $n$  point rule in one dimension can be used to integrate a polynomial of order  $2n + 1$  exactly (Stroud 1971; Conte & de Boor 1980). So a  $3 \times 3 \times 3$  rule is required to integrate the cubic interpolation function in the 8-node element exactly. But it has been observed that in practice the reduced integrated element converges faster to the exact solution. This technique which is widely used in the FEM is called reduced integration. These rules need to be used with care. A very low order rule can lead to mechanisms (Brassioulis 1989), while the use of a very high order one leads to delayed convergence. One common type of mechanism encountered during reduced integration is the presence of hour glass modes. Various methods have been proposed in order to control these modes for 8 noded hexahedral solid elements (Flagan & Belytschko 1981; Schulz 1985) and for 20-noded solid elements (Kelen 1989). Such rules have proved their mettle in many problems, particularly related to plate and shell flexure (Pawsey & Clough 1971; Zienkiewicz *et al* 1971) and also when solid elements are degenerated into flat plate and curved shell elements. Explicit integration for such degenerated elements have led to difficulties, which require the use of special techniques (Vlachoutsis 1990). The success of this techniques has led to efforts to include the experience gained into updating finite element codes (Case & Vandegrift 1986; Case *et al* 1986). Many explanations have been sought for the success. Some are heuristic. Others have been based on more scientific arguments. Prathap and co-workers (1992) have looked upon this technique as a method to obtain "field consistency". Zienkiewicz & Taylor (1989) attribute the success to the fact that the Gauss points being used in reduced integration are exactly the optimal points for stress recovery. "However, the main reason for success does not lie here but is associated with the fact that it provides the necessary singularity of the constraint part of the matrix, which avoids locking" (Zienkiewicz &

Taylor 1989). It should also be noted here that for mapped elements (like the isoparametric ones) full integration will not exactly integrate a deformed element (Cook *et al* 1989).

(2) Addition of bubble modes (Wilson *et al* 1973; Wilson 1973): This technique involves the addition of certain degrees of freedom not associated with any node (like the hierarchic functions mentioned previously). This brought into use, incompatible elements, where the displacement fields are not continuous across element boundaries. The variables associated with the nodeless degrees of freedom are later condensed out using a static condensation procedure.

In the case of the 8-node brick element this technique when used in tandem with reduced integration, gives very good results and has found its way into many commercial finite element programs. Unfortunately, the use of this technique requires experience. While it is possible that certain polynomial terms can be associated with the element whose "shape functions" are known, no clues are available in literature to guide the novice as to which functions are to be or not to be chosen. The success of the technique appears to lie in the fact that the functions which are chosen are the exact ones required to remove only certain types of locking. For example, in the case of the 8-node brick element, the incompatible modes chosen alleviate parasitic shear stored during bending. It also alleviates locking near incompressible limits.

(3) Using unequal order interpolation: This is one of the simpler techniques in use, especially in the case of 1- and 2-dimensional elements. For example, Tessler & Dong (1981) have formulated one such Timoshenko beam element. Here the order of the interpolation function used for the rotational degrees of freedom is one less than that of the translational ones. Its success could be attributed to the fact that the terms dropped from the interpolation functions for the rotational degrees of freedom are exactly the ones which if present will cause locking. Unequal order interpolation has been used in the formulation of many solid finite elements, like transitional elements and solid elements in the analysis of plates (16-node hexahedral elements).

(4) Assumed strain methods (MacNeal 1982; Olesan 1983): This techniques involves the use of computation of interpolation functions (of lower order) and smoothing them in some least square sense. It has been felt that this procedure is equivalent to that of reduced integration (Prathap 1992), but it has an added advantage in that the procedure can be used to obtain interpolation functions, while in the case of reduced integration, the points of reduced integration may not exist.

(5) Residual energy balancing (Fried 1974, 1975; Cook 1977): In this technique certain constraints contributing to locking are identified. These constraints are then artificially removed by the use of a constant, which the designer of the finite element sets to an arbitrarily small value. The point to be noted here is that the arbitrary constant to be used is problem dependent and it appears to be difficult to choose one value for a set of elements and problems. The constant is also mesh dependent, thereby compounding to the confusion. Stresses predicted by this method are "very unreliable" (Prathap 1992) and grossly dependent on the value of the scaling constant chosen. This technique, it is felt, is extrapolatable to solid finite elements and hence must be used with care.

(6) Reduced interaction (Prathap 1992): In this case the field variable that causes locking is identified and replaced by a lower order interpolation field.

(7) Elements based on variational principles other than the principle of minimum total potential: These will be dealt with in a later section.

There are a number of other ingenious techniques used to obtain better elements. Amongst the others are synthesis using Fourier components (Park 1984), the use of trigonometric interpolation functions (Heppler & Hansen 1987), use of shear constraints (Crisfield 1984) etc. The list appears to grow even today, as newer techniques are developed to obtain elements which would predict accurate solutions (for both stresses and displacements). Obviously, the primary reason for the development of these elements could be attributed to the lack of success with those previously developed. Even the techniques mentioned above do not appear to provide satisfactory explanations for their success. While, sometimes they work well in one context, they do not do so in others. This has led to a compounding of the already prevailing confusion.

#### **4. Some other types of interpolation functions**

In addition to the interpolation functions mentioned above, various other types are still being developed. The hybrid-stress elements (see § 5 below) which uses independent interpolations for stresses and displacements has been viewed as just another extra-variational technique by some authors (Prathap 1992; Andelfinger & Ramm 1993). Some authors have looked at the use of "rational" interpolation functions (Wachspress 1975) for formulating a hexahedral element (Wait 1971) and its applications (Yu 1990). Here interpolation functions which can be defined as the ratio of two polynomials are used. De Freitas & Castro (1992) have proposed the use of digital interpolation functions (based on a class of functions known as the Walsh functions) and can be viewed as a method similar to the ones using trigonometric functions above. A new class of interpolation functions called "physical shape functions" have also been presented (Gilewski & Gomulinski 1990), where physical parameters and material parameters are interpolated in addition to displacements. Bergan & Nygard (1984) propose a "free formulation" where elements can be formulated by splitting up the displacement field into one with lower order fields and another with higher order ones. Even though it appears that these techniques have not yet been used in the design and development of solid finite elements, they can, in some sense, be described as generalized methods, and we feel they are extrapolatable to solid elements.

#### **5. The newer 8-node brick elements**

It has been seen that it is the lower-order hexahedral elements (in the case of two-dimensional elements it is the quadrilateral ones) which suffer from locking due to parasitic shear. Therefore it is not surprising that recent developments of solid finite elements have addressed themselves to this problem. In this section we look at some "new" formulations of the 8-node brick element.

##### **5.1 Displacement-based elements**

We first look at the displacement-based elements (like all formulations discussed above). A number of such elements have been formulated. Wilson (1973) first proposed an incompatible element, with additional bubble modes in order to incorporate a method which ensured that the 8-node brick does not lock under bending situations.

The element he proposed did not pass the patch test. Taylor *et al* (1976) by imposing a set of constraints which ensured that the element satisfies the patch test *a priori*, were able to arrive at an element which has remained popular to this day. Bretl & Cook (1979) have proposed an 8-node element, in which the stiffness matrix corresponding to the lower-order strains and those of the higher-order ones have been separated out. Depending on the choice of the problem, the corresponding higher-order strains are used. The stiffness matrices of the lower-order fields use a stress-based approach different from the higher-order ones. Yunus *et al* (1991) have proposed an 8-node element with rotational degrees of freedom. This element which had three additional modes to overcome Poisson ratio locking, is prone to experience zero energy modes which are alleviated using additional procedures. (The method is similarly extended to the development of tetrahedral elements with rotational degrees of freedom (Pawlak *et al* 1991). All the displacement-based elements discussed so far predict poor stresses under many loading conditions. Wilson & Ibrahimbegovic (1990) using a least squares approximation to extract stresses were able to improve the stress prediction of the 8-node brick element with incompatible modes – stress predictions were identical to that predicted by stress-based elements. In this way they were able to separate the procedures for computing the stresses and for displacements.

## 5.2 Elements based on other variational principles

The discussions above highlight the difficulties encountered by designers in the formulation of displacement based elements. Over the years elements based on other variational principles have been developed.

Prathap and co-workers (1992) have attempted to obtain explanations for the problem of locking and to formulate elements which do not take recourse to any of the extra-variational techniques. They use the “field consistency paradigm” and the “functional reconstitution technique” to obtain elements which they show do not lock. This method appears to call for a variation of the strain fields independent of the displacements and is based on the more general Hellinger–Reissner and the Hu–Washizu principles (Washizu 1968). Moreover the solid elements which have been formulated using this technique suffer from one severe drawback in that the 8-node brick formulated (Chandra & Prathap 1989) is very sensitive to geometric distortions and predicts poor stress values (Dong & de Freitas 1992). This has been attributed to the difficulties encountered in finding a “technique of consistent mapping for covariant finite element formulation in 3-D applications...” (Naganarayana 1991). In some cases (Naganarayana & Prathap 1991), reduced/selective integration is used in order to integrate the strain energy to overcome this difficulty and is illustrated with a 27-node hexahedral brick element.

Amongst the more popular formulations are the ones in which compatible displacements and equilibrating stresses are independently interpolated. Stress parameters are eliminated at the element level and a stiffness matrix is obtained, as by Pian (1973). These formulations are called hybrid/mixed formulations. Many three-dimensional hybrid/mixed stress elements have been developed (Zienkiewicz & Taylor 1989). Here also extravariational techniques like reduced integration and bubble modes are used. Other elements formulated using these principles are 8-node elements (Irons 1972; Lee 1974), 20-node elements (Ahmad & Irons 1974), special purpose three-dimensional

elements for thick plate analysis (Spilker 1981) and 20-node quadratic displacement 3-dimensional isoparametric hybrid elements (Spilker & Singh 1982).

Using this method in which displacement variables are separated, Pian and co-workers were able to formulate isoparametric elements which are coordinate invariant. These elements are known to have certain drawbacks. One of them is the presence of zero energy modes which appear when stress interpolants are used. Pian & Chen (1983) were able to establish a method which prevented the appearance of zero energy deformation modes. This concept was extended by Pian & Tong (1986) to develop hexahedral hybrid elements. Spilker & Singh (1982) developed a 20-node hybrid hexahedral element based on the complementary energy principle. As the stress interpolants need to satisfy equilibrium conditions, an *a priori* cartesian coordinate interpolation was necessary. Such interpolation is known to provide poor stresses and displacements under grossly distorted conditions. Another such element was formulated by Loikkanen & Irons (1984). Punch & Atluri (1984a, 1984b) used symmetric group theory to study a number of such elements (using cartesian interpolation functions) for stability, coordinate invariance and for the optimal stress functions to be chosen. Tang & Chen (1982) proposed a series of non-conforming stress based elements. Even though their performance was better than those of conforming ones, they were no superior to the modified Wilson 8-node brick. Chen & Cheung (1987) derived a new functional (as a generalization of the functional proposed by Pian in 1964, with displacements, stresses and strains as independent variables) to obtain a series of isoparametric elements (Cheung & Chen 1988). However, the formulation of these elements requires an additional stress field and proper matching of displacements, strains and stresses. This requires experience.

Bachrach (1987) used a  $\gamma$ -projection operator in tandem with the Hu–Washizu principle to obtain a stress-based element which performed well in bending and near incompressibility. The element required to be reduced integrated and stresses extrapolated to the nodes. Yunus *et al* (1989) have formulated a hybrid solid element with rotational degrees of freedom. Sze *et al* (1990) by using a modified Hellinger–Reissner principle, with prior constraints and recognition of the orthogonality of strains, stresses and incompatible displacements obtain a series of isoparametric elements. These could be reduced integrated. Chen & Cheung (1992) by using a weaker constraint condition on their previous formulation (Cheung & Chen 1988) were able to obtain elements with improved performance. Dong & de Freitas (1992) obtained an isoparametric incompatible solid element based on the Hellinger–Reissner principle which was modified by the presence of a constant stress multiplier. Sze & Ghali (1993) start with the assumed stress element of Pian & Tong (1986), identify the strain components which cause locking and selectively scale them down to obtain an incompatible element. The element is then corrected to pass the patch test using an “admissible matrix formulation”. Andelfinger & Ramm (1993) use the enhanced assumed strain method (EAS) (obtained by the use of an extra strain field – not continuous) to obtain an 8-node element. Pian (1985) by a modification of his original procedure, by choosing the displacement field in a “consistent” manner, was able to obtain a hybrid 8-node solid element with improved performance under distorted conditions. The growth in popularity and reliability of current day symbolic manipulators has speeded up the development of solid elements. Tan *et al* (1991) have elaborated on procedures to formulate an 8-node hybrid hexahedral element.

All the elements mentioned above have been compared individually against the displacement based solid elements. Comparisons are normally made against standard



problems proposed in literature. (MacNeal & Harder 1985; Belytschko & Liu 1986; White & Abel 1989). Kim *et al* (1990) have compared the performance of the degenerated 20-node element against the 18-node assumed strain element.

All the solid elements discussed do not yet seem to show completely correct predictions for stresses as well as displacements encompassing different types of loading conditions.

## 6. Elements based on solutions to the Navier equation

All the elements discussed in the preceding sections seem to be formulated with the express desire of obtaining good element performance. Instead of attacking the root of the problem, a precise statement of which does not seem to exist so far, they seem to provide ways and means of getting around it. Even methods like the field consistency concept and the use of physical shape functions, which attempt to get to the bottom of the problem, do not as yet appear to be perfected for solid finite elements.

We have looked at the problem afresh. Recognizing the fact that the overall element behaviour seems to be guided by the interpolation functions chosen, we have formulated two hexahedral elements (Venkatesh & Shrinivasa 1993a) using Papcovitch-Neuber solutions as interpolation functions which satisfy the governing differential equations (Navier equations) *a priori* (Venkatesh & Shrinivasa 1993b). It was felt that such a method would combine the simplicity of the displacement-based approach (the Navier equations are written in terms of displacements) and accurate stress prediction of the stress-based approach (the Navier equations are based on equations of stress equilibrium). Such a method has been thought of before (Hoppe 1973), but does not seem to have been implemented for solid elements.

These elements (Venkatesh & Shrinivasa 1993c, 1993d) have been formulated without taking recourse to any of the above mentioned extra-variational techniques. The elements are exactly integrated by breaking the hexahedra into five tetrahedra (Stroud 1971) and then using appropriate gauss point integration. They exhibit correct element behaviour when subjected to bending moment loads even under grossly distorted conditions and also when the stresses are computed directly by evaluating them at the nodes instead of "at some optimal locations within the elements" (Barlow 1976, 1989). The element performance for out-of-plane shear matches with that of the Wilson 8-node brick for rectangular elements and has been shown to be better than the latter under distorted conditions. The elements pass the constant strain patch test under rectangular conditions. Under distorted conditions they pass the "weak patch test" (Cook *et al* 1989). Relative performance of the these elements *vis-a-vis* already existing ones can be found in Venkatesh & Shrinivasa (1993e).

It therefore appears that it would be possible to formulate finite elements which do not experience any irrational behaviour by simply using the solutions to the Navier equations as interpolation functions.

## 7. Conclusion

We have reviewed the development of solid finite elements since the inception of the finite element method. We have looked at displacement based elements and their drawbacks such as the presence of shear locking etc. We have briefly reviewed the

extra-variational techniques used in literature to overcome these drawbacks. The paper also reviews the state-of-the-art lower-order hexahedral finite elements being currently proposed. Finally, we critically discuss the lower-order finite elements developed by using interpolation functions which satisfy the governing differential equations *a priori* and conclude that it is possible to formulate such finite elements without taking recourse to any extra-variational techniques.

The authors wish to acknowledge the support received for carrying out this work from the Aeronautics Research and Development Board through grant no. AERO/RD-134/100/10/90-91/653.

## References

- Ahmad S, Irons B M 1974 An assumed stress approach to refined isoparametric finite elements in three dimensions. In *Finite element methods in engineering*, University of New South Wales
- Ahmad S, Irons B M, Zienkiewicz O C 1970 Analysis of thick and thin shell structures by curved elements. *Int. J. Numer. Methods Eng.* 2: 419-451
- Andelfinger U, Ramm E 1993 EAS-elements for two-dimensional, three-dimensional, plate and shell structures and their equivalence to HR-elements. *Int. J. Numer. Methods Eng.* 36: 1311-1337
- Argyris J H 1965 Matrix analysis of three-dimensional elastic media small and large displacements. *AIAA J.* 3: 45-51
- Argyris J H, Fried I 1968a The LUMINA element for the matrix displacement method. *Aeronaut. J. R. Aeronaut. Soc.* 72: 514-517
- Argyris J H, Fried I, Scharpf D W 1968b The HERMES8 element for the matrix displacement method. *Aeronaut. J. R. Aeronaut. Soc.* 72: 613-617
- Argyris J H, Fried I, Scharpf D W 1968c The TET20 and TEA8 elements for the matrix displacement method. *Aeronaut. J. R. Aeronaut. Soc.* 72: 618-623
- Bachrach W E 1987 An efficient formulation of hexahedral elements with high accuracy for bending and incompressibility. *Comput. Struct.* 26: 45-467
- Barlow J 1976 Optimal stress locations in finite element models. *Int. J. Numer. Methods Eng.* 10: 243-251
- Barlow J 1989 More on optimal stress points-reduced integration, element distortions and error estimation. *Int. J. Numer. Methods Eng.* 28: 1487-1504
- Bathe K, Wilson E L 1976 *Numerical methods in finite element analysis* (Englewood Cliffs, NJ: Prentice-Hall)
- Bathe K J, Dvorkin E N 1985 A four-node plate bending element based on the Mindlin/Reissner plate theory and mixed interpolation. *Int. J. Numer. Methods Eng.* 21: 367-383
- Belytschko T, Liu W K 1986 Test problems and anomalies in shell finite elements. In *Reliability of methods for engineering analysis Proc. Int. Conf. Univ. College* (Swansea: UK: Pineridge Press) pp 393-406
- Bergan P G, Nygard M K 1984 Finite elements with increased freedom in choosing shape functions. *Int. J. Numer. Methods Eng.* 20: 643-663
- Brassioulis D 1989 On the basics of the shear locking problem of  $C^0$  isoparametric plate elements. *Comput. Struct.* 33: 169-185
- Bretl J L, Cook R D 1979 A new eight-node solid element. *Int. J. Numer. Methods Eng.* 14: 593-615
- Case W R, Mitchell R S, Vandegriff R E 1986 Improved accuracy in solid isoparametric elements using selectively reduced integration. *NASA Conference Publication 2373*, NASA, Washington, DC, p. 31
- Case W R, Vandegriff R E 1986 Improved isoparametric solid and membrane elements. *NASA Conference Publication 2419*, NASA, Washington, DC, pp. 28-55

- Chacour S 1972 'DANUTA', a three-dimensional finite element program used in the analysis of turbomachinery. *J. Basic Eng.*, ASME 94: 71-77
- Chandra S, Prathap G 1989 A field-consistent formulation for the 8-noded solid finite element. *Comput. Struct.* 33: 345-355
- Chen W, Cheung Y K 1987 A new approach for the hybrid element method. *Int. J. Numer. Methods Eng.* 24: 1697-1709
- Chen W, Cheung Y K 1992 Three-dimensional 8-node and 20-node refined hybrid isoparametric elements. *Int. J. Numer. Methods Eng.* 35: 1871-1889
- Cheung Y K, Chen W 1988 Isoparametric hybrid hexahedral elements for three dimensional stress analysis. *Int. J. Numer. Methods Eng.* 26: 677-693
- Ciarlet P G 1978 *The finite element method for elliptic problems* (New York: North Holland)
- Clough R W 1969 Comparison of three-dimensional finite elements In *Symposium on application of finite element methods in civil engineering* (eds) W H Rowan, R M Hackett (Nashville, TN: ASCE)
- Collatz L 1950 *The numerical treatment of differential equations* (Berlin: Springer-Verlag)
- Conte S D, de Boor C 1980 *An introduction to numerical analysis, an algorithmic approach* (Singapore: McGraw Hill)
- Cook R D 1977 More about artificial softening of finite elements. *Int. J. Numer. Methods Eng.* 11: 1334-1339
- Cook R D, Malkus D S, Plesha M E 1989 *Concepts and applications of finite element analysis* (New York: John Wiley and Sons)
- Courant R 1943 Variational methods for the solution of problems of equilibriums and vibrations. *Bull. Am. Math. Soc.* 49: 1-23
- Courant R, Hilbert D 1953 *Methods of mathematical physics* (New York: John Wiley and Sons)
- Crisfield M A 1984 A quadratic mindilin element using shear constraints. *Comput. Struct.* 18: 833-852
- de Arrantes Oliveira E R 1968 Theoretical foundations of the finite element method. *Int. J. Solids Struct.* 4: 929-952
- de Freitas J A T, Castro L M S 1992 Digital interpolation in mixed finite element structural analysis. *Comput. Struct.* 44: 743-751
- Donea J, Lamain L G 1987 A modified representation of transverse shear  $C^0$  quadrilateral plate elements. *Comput. Methods. Appl. Mech. Eng.* 63: 183-207
- Dong Y F, de Freitas J A T 1992 An efficient 8-node incompatible solid element with stress interpolation. *Comput. Struct.* 44: 773-781
- Ergatoudis J, Irons B M, Zienkiewicz O C 1968 Three-dimensional analysis of arch dams and their foundations. In *Symposium on arch dams* (London: Institution of Civil Engineers)
- Ferguson G H, Clark R D 1979 A variable thickness, curved beam and shell stiffening element with shear deformations. *Int. J. Numer. Methods. Eng.* 14: 581-592
- Fjeld S A three-dimensional theory of elasticity. In *Finite element methods in stress analysis* (eds) I Holland, K Bell (Trondheim: Tapir Norway)
- Flagan D P, Belytschko T 1981 A uniform strain hexahedron and quadrilateral with orthogonal hour glass control. *Int. J. Numer. Methods. Eng.* 17: 679-706
- Fried I 1974 Residual energy balancing technique in the generation of plate bending elements. *Comput. Struct.* 8: 771-778
- Fried I 1975 Finite element analysis of thin elastic shells with residual energy balancing and the role of rigid body modes. *J. Appl. Mech.* 42: 99-104
- Gallagher R H, Padlog J, Bijlard P P 1962 Stress analysis of heated complex shapes. *J. Am. Rocket Soc.* 32: 700-707
- Gilewski W, Gomulinski A 1990 Physical shape functions: a new concept in finite elements. *Finite Element News* (3): 20-23
- Haggenmacher G W 1993 The shape of things. III. *Finite Element News* (2): 38-41
- Heppler G R, Hansen J S 1987 Timoshenko beam finite elements using trigonometric basis functions. *AIAA J.* 26: 1378-1386
- Hoppe V 1973 Finite elements with harmonic interpolation functions. In *The mathematics of finite elements and applications* (ed.) J R Whiteman (New York: Academic Press)
- Hughes J R, Allik H 1969 Finite elements for compressible and incompressible continua. In *Symposium on application of finite element methods in civil engineering* (ed.) W H Rowan, R M Hackett (Nashville, TN: ASCE)

- Irons B M 1966 Engineering applications of numerical integration in stiffness methods. *AIAA J.* 4: 2035-2037
- Irons B M 1972 An assumed stress version of the Wilson 8-node brick. University of Wales, Computer Report CNME/CR/56
- Irons B M, Ahmad S 1979 *Techniques of finite elements* (Chichester: Ellis Horwood)
- Kelen P 1989 Control of spurious nodes in 20-noded isoparametric element. *Commun. Appl. Numer. Methods* 5: 415-422
- Kidger D J 1990 The 14-node brick element family. *Finite Elements Anal. Design* 6:
- Kim Y H, Jones R F, Lee S W 1990 Study of 20-node solid element. *Commun. Appl. Numer. Methods* 6: 197-205
- Lee S W 1974 An assumed stress hybrid finite element for three-dimensional elastic structural analysis. *Massachusetts Institute of Technology*, ASRL-TR-170-3
- Loikkanen M J, Irons B M 1984 An 8-node brick finite element. *Int. J. Numer. Methods. Eng.* 20: 523-528
- MacNeal R H 1982 Derivation of assumed strain matrices by assumed strain distributions. *Nucl. Eng. Design* 70: 3-12
- MacNeal R H, Harder R L 1985 A proposed set of problems to test finite element accuracy. *Finite Elements Anal. Design* 1: 3-20
- MacNeal R H 1992 On the limits of finite element perfectibility. *Int. J. Numer. Methods. Eng.* 35: 1589-1601
- Melosh R J 1963 Structural analysis of solids. *J. Struct. Div. ASCE* 89: 205-223
- Naganarayana B P 1991 *Consistency and correctness principles in quadratic displacement type finite elements*. Ph D thesis, Dept. Aerospace Eng., Indian Institute of Science, Bangalore
- Naganarayana B P, Prathap G 1991 Field-consistency analysis of 27-noded hexahedral elements for constrained media elasticity. *Finite Elements Anal. Design* 9: 149-168
- Nguyen V P 1982 Automatic mesh generation with tetrahedron elements. *Int. J. Numer. Methods. Eng.* 18: 273-289
- Oden J T 1983 *Finite elements - mathematical aspects* (Englewood Cliffs, NJ: Prentice-Hall)
- Oleson J F 1983 Field redistribution in finite elements - a mathematical alternative to reduced integration. *Comput. Struct.* 17: 157-159
- Pammer Z, Szabo L 1981 Stereo decomposition subroutines for three-dimensional plotter programs. *Int. J. Numer. Methods. Eng.* 17: 1571-1575
- Park K C 1984 Symbolic fourier analysis procedures for  $C^0$  finite elements. In *Innovative Methods Nonlinear Anal.* (Swansea, UK: Pineridge Press) 269-293
- Pawlak T P, Yunus S M, Cook R D 1991 Solid elements with rotational degrees of freedom. Part II. Tetrahedron elements. *Int. J. Numer. Methods. Eng.* 31: 593-610
- Pawsey S F, Clough R W 1971 Improved numerical integration of thick shell finite elements. *Int. J. Numer. Methods. Eng.* 3: 575-586
- Peano A G 1976 Hierarchics of conforming finite elements for elasticity and plate bending. *Comput. Math. Appl.* 2: 3-4
- Pian T H H 1973 Hybrid models. In *Numerical methods and computer methods in applied mechanics* (eds) S J Fenves, N Perrone, A R Robinson, W C Schnobrich (New York: Academic Press)
- Pian T H H 1985 Finite elements based on consistently assumed stresses and displacements. *Finite Elements Anal. Design* 1: 131-140
- Pian T H H, Chen D P 1983 On the suppression of zero energy deformation modes. *Int. J. Numer. Methods. Eng.* 19: 1741-1752
- Pian T H H, Tong P 1986 Relations between incompatible displacement model and hybrid stress model. *Int. J. Numer. Methods. Eng.* 22: 173-181
- Prathap G 1985 A  $C^0$  continuous 4-noded cylindrical shell element. *Comput. Struct.* 21: 995-999
- Prathap G 1992 Recent advances in finite element technology. NAL-UNI Series, National Aerospace Lab., Bangalore
- Punch E F, Atluri S N 1984a Applications of isoparametric three-dimensional hybrid-stress finite elements with least-order fields. *Comput. Struct.* 19: 406-430
- Punch E F, Atluri S N 1984b Development and testing of 2- and 3-d hybrid stress elements. *Comput. Methods. Appl. Mech. Eng.* 47: 331-356
- Rashid Y R 1969; 1970 Three dimensional analysis of elastic solids: I. Analysis procedure; II. The computational problems. *Int. J. Solids Struct.* 5: 6: 1311-1331; 195-207

- Rashid Y R, Smith P D, Prince N 1969 On further application of finite element method to three-dimensional elastic analysis. In *Symposium on high speed computing of elastic structures* (eds) W H Rowan, R M Hackett (Nashville, TN: ASCE)
- Rigby G L, McNeice G M 1972 A strain energy basis for studies of element stiffness matrices. *AIAA J.* 10: 1490–1493
- Robinson J 1985 *Early FEM pioneers* (Dorset: Robinson and Associates)
- Schulz J C 1985 Finite element hourglassing control. *Int. J. Numer. Methods. Eng.* 21: 1039–1048
- Smith I M, Kidger D J 1992 Elastoplastic analysis using the 14-node brick element family. *Int. J. Numer. Methods. Eng.* 1263–1275
- Spilker R L 1981 High-order three-dimensional hybrid-stress elements for thick plate analysis. *Int. J. Numer. Methods. Eng.* 17: 53–69
- Spilker R L, Singh S P 1982 Three-dimensional hybrid-stress isoparametric quadratic displacement elements. *Int. J. Numer. Methods. Eng.* 18: 445–465
- Stolarski H, Belytschko T 1981 Membrane locking and reduced integration for curved beam elements. *J. Appl. Mech.* 49: 172–178
- Strang G 1972 Variational crimes in the finite element method. In *The mathematical foundations of the finite element method with applications to partial differential equations* (ed.) A K Aziz (New York: Academic Press)
- Strang G, Fix G J 1973 *An analysis of the finite element method* (Englewood Cliffs, NJ: Prentice-Hall)
- Stroud A H 1971 *Approximate calculation of multiple integrals* (Englewood-Cliffs, NJ: Prentice-Hall)
- Synge J L 1957 *The hypocircle in mathematical physics* (London: Cambridge University Press)
- Sze K Y, Chow C L, Chen W 1990 A rational formulation of isoparametric hybrid stress elements. *Finite Elements Anal. Design* 7: 61–72
- Sze K Y, Ghali A 1993 Hybrid hexahedral element for solids, plates, shells and beams by selective scaling. *Int. J. Numer. Methods. Eng.* 36: 1519–1540
- Tan H Q, Chang T Y P, Zheng D 1991 On symbolic manipulation and code generation of a hybrid three-dimensional solid element. *Eng. Comput.* 7: 47–59
- Tang L, Chen W 1982 The non-conforming finite elements for stress analysis. In *Proceedings in the International Conference on finite element methods* (eds) G He, Y K Cheung (New York: Gordon and Breach)
- Taylor R L, Beresford P J, Wilson E L 1976 A non-conforming element for stress analysis. *Int. J. Numer. Methods. Eng.* 10: 1211–1219
- Tessler A, Dong S B 1981 On a hierarchy of conforming beam elements. *Comput. Struct.* 14: 335–344
- Turner M J, Clough R W, Martin H C, Topp L J 1956 Stiffness and deflection analysis of complex structures. *J. Aeronaut. Sci.* 23: 805–823
- Venkatesh D N, Shrinivasa U 1993a Formulation of finite elements which satisfy the Navier equation *a priori*. Report 93 VAR5, Dept. Mech. Eng. Indian Institute of Science, Bangalore
- Venkatesh D N, Shrinivasa U 1993b Solutions to the Navier equation and equivalent Papcovitch-Neuber functions. Report 93 VAR1, Dept. Mech. Eng. Indian Institute of Science, Bangalore
- Venkatesh D N, Shrinivasa U 1993c Generation of an 8-node brick element using PN functions. Report 93 VAR2, Dept. Mech. Eng. Indian Institute of Science, Bangalore
- Venkatesh D N, Shrinivasa U 1993d Hexahedral elements using PN functions for beam problems. Report 93 VAR3, Dept. Mech. Eng. Indian Institute of Science, Bangalore
- Venkatesh D N, Shrinivasa U 1993e Performance study of hexahedral elements using PN functions. Report 93 VAR5, Dept. Mech. Eng. Indian Institute of Science, Bangalore
- Vlachoutsis S 1990 Explicit integration of three-dimensional degenerated shell elements. *Int. J. Numer. Methods. Eng.* 29: 861–880
- Wachspress E I 1975 *Rational finite element basis* (New York: Academic Press)
- Wait R 1971 A finite element for 3-dimensional function approximation. *Proc. Conf. Appl. Numer. Anal. Springer-Verlag lecture notes in Mathematics* (Berlin, Heidelberg: Springer-Verlag) 228: 348–352
- Washizu K 1968 *Variational principles in elasticity and plasticity* (London: Pergamon)
- White D W, Abel J F 1989 Testing of shell finite element accuracy and robustness. *Finite Elements Anal. Design* 6: 129–151

- Wilson E L 1973 Incompatible displacement models, In *Numerical methods and computer methods in applied mechanics* (eds) S J Fenves, N Perrone, A R Robinson, W C Schnobrich (New York: Academic Press)
- Wilson E L, Ibrahimbegovic A 1990 Use of incompatible displacement modes for the calculation of element stiffnesses or stresses. *Finite Elements Anal. Design* 7: 229–241
- Wilson E L, Taylor R L, Doherty W P, Ghaboussi T 1973 Incompatible displacement models. In *Numerical Computational Methods in Structural Mechanics* (eds) S J Fenves *et al* (New York: Academic Press)
- Yang T Y 1986 *Finite element structural analysis* (Englewood Cliffs, NJ: Prentice-Hall)
- Yu H S 1990 A rational displacement interpolation function for axisymmetric finite element analysis of nearly incompressible materials. *Finite Elements Anal. Design* 10: 205–219
- Yunus S M, Pawlak T P, Cook R D 1991 Solid elements with rotational degrees of freedom: Part I. Hexahedron elements. *Int. J. Numer. Methods. Eng.* 31: 573–592
- Yunus S M, Saigal S, Cook R D 1989 On improved hybrid elements with rotational degrees of freedom. *Int. J. Numer. Methods. Eng.* 28: 785–800
- Zienkiewicz O C 1973 Finite elements – the background story. In *The mathematics of finite elements and applications* (ed.) J R Whiteman (New York: Academic Press)
- Zienkiewicz O C 1977 *The finite element method* (New York: McGraw Hill)
- Zienkiewicz O C, De J P, Gago S R, Kelly D W 1983 The hierarchical concept in finite element analysis. *Comput Struct.* 16: 53–65
- Zienkiewicz O C, Irons B M, Ergatoudis J, Ahmad S, Scott F C 1969 Isoparametric and associated element families for two and three dimensional analysis. In *Finite element methods in stress analysis* (eds) I Holland, K Bell (Trondheim, Norway: Tapir)
- Zienkiewicz O C, Irons B M, Scott F C, Campbell J S 1970 Three dimensional stress analysis. In *Symposium on high speed computing of elastic structures* 1: 413–432
- Zienkiewicz O C, Taylor R L 1989 *The finite element method. Vol. I: Basic formulation and linear problems* (New York: McGraw Hill)
- Zienkiewicz O C, Taylor R L, Too J M 1971 Reduced integration technique in general analysis of plates and shells. *Int. J. Numer. Methods. Eng.* 3: 275–290



## Development of robust finite elements for general purpose structural analysis

G PRATHAP\*, B P NAGANARAYANA and B R SOMASHEKAR

Structural Sciences Division, National Aerospace Laboratories, Bangalore, 560 017, India

\* Also, Jawaharlal Nehru Centre for Advanced Scientific Research, Indian Institute of Science Campus, Bangalore 560 012, India

MS received 20 August 1993; revised 18 February 1994

**Abstract.** The finite element method emerged out of the old work and energy methods and matrix structural analysis to become a numerical procedure to solve practical stress analysis problems in solid and structural mechanics. With the impetus given by the rapid development of computer technology, it became the most overwhelmingly popular analysis and design computational tool for a very wide spectrum of engineering science, e.g. fluid mechanics, heat transfer and electro-magnetics. Today, there are very powerful general-purpose software codes that make analyses and design tasks that were once considered to be intractable, routinely simple. Many of these are closely held proprietary codes owned and used in-house by large engineering firms or sold or licensed and supported by specialist companies. (Recent estimates indicate that the market for these codes has reached a turnover of a billion dollars and that industries and institutions spend several tens of billions of dollars in running such codes.) These codes are rarely given out in source code. In order to have an in-house code that could be continuously up-graded and enhanced, NAL initiated some work to develop a medium-sized general purpose code (about 20,000 lines of FORTRAN code) for the analysis of laminated composite structures (FEPACS – finite element package for analysis of composite structures), recognising the importance that laminated composites were assuming in aerospace structural technology.

Several key elements commonly found in general purpose packages (GPP) used by the aerospace, automobile and mechanical engineering industries were identified. These were re-designed incorporating anisotropic composite capabilities and validated. Many hurdles were faced during this task and required an examination of the basic issues at a paradigmatic level. Concepts such as consistency and variational correctness were introduced and studied critically. These guidelines played a critical role in developing robust versions of the elements and are briefly covered in this review. The paradigms also helped to identify procedures to perfr



*a priori* error estimates for the quality of approximation and this allowed the elements being developed to be critically validated.

The article concludes with a summary of what has been achieved and also suggests areas where the concepts can be applied fruitfully in the study of the displacement type finite element method.

**Keywords.** Finite element method; composite structures; structural analysis; element technology; general purpose packages; FEPACS.

## 1. Introduction – from $C^1$ to $C^0$ elements

We shall examine the subject as seen from our own practical viewpoint and as it must have been seen by the larger finite element community as well, as we undertook the task of designing a library of simple, accurate and efficient elements for general purpose finite element structural analysis.

At the time we began our work, around 1978, it was slowly becoming accepted that the element libraries of major general purpose packages (GPP) were replacing what were called the  $C^1$  elements with what were known as the  $C^0$  elements. The former were based on well-known classical theories of beams, plates and shells (i.e. the Euler–Bernoulli beam theory, the Kirchhoff–Love plate theory and the equivalent shell theories), reflecting the confidence that structural analysts have had in such theories for over two centuries. It is indicative that the early history of the finite element technology was almost entirely confined to the use of elements based on such theories. These theories did not allow for transverse shear strain and permitted the modelling of such structures by defining deformation in terms of a single field,  $w$ , the transverse deflection of a point on what is called the neutral axis (in a beam) and neutral surface of a plate or shell. The strains could then be computed quite simply from the assumption that normals to the neutral surface remained normal after deformation. One single governing differential equation resulted, although of a higher order (in comparison to other theories we shall discuss shortly), and this was considered to be an advantage.

There were some consequences arising from such an assumption both for the mathematical modelling aspect as well as for the finite element (discretisation) aspect. In the former, it turned out that to capture the physics of deformation of thick or moderately thick structures, or the behaviour of plates and shells made of newly emerging materials such as high performance laminated composites, it was necessary to turn to more general theories accounting for transverse shear deformation as well – these required the definition of rotations of normals which were different from the slopes of the neutral surface. Some of the contradictions that arose as a result of the old  $C^1$  theories – e.g. the use of the fiction of the Kirchhoff effective shear reactions could now be removed, restoring the more physically meaningful set of three boundary conditions on the edge of a plate or shell (the Poisson boundary conditions as they are called) to be used. The orders of the governing equations were correspondingly reduced. A salutary effect that carried over to finite element modelling was that the elements could be designed to have nodal degrees of freedom which were the six basic engineering degrees of freedom – the three translations and the three rotations at a point. This was ideal from the point of view of the organization of a general purpose package. Also, elements needed only simple basis functions requiring only

the continuity of the fields across element boundaries – these are called the  $C^0$  requirements. In the older  $C^1$  formulations, continuity of slope was also required and to achieve this in arbitrarily oriented edges, as would be found in triangular or quadrilateral planforms of a plate bending element, it was necessary to retain curvature degrees of freedom ( $w_{xx}, w_{xy}, w_{yy}$ ) at the nodes and rely on quintic polynomials for the element shape or basis functions. So, as general purpose packages ideal for production run analyses and design increasingly found favour in industry, the  $C^0$  beam, plate and shell elements slowly began to replace the older  $C^1$  equivalents. It may be instructive to note that the general two-dimensional (i.e. plane stress, plane strain and axisymmetric) elements and three-dimensional (solid or brick as they are called) elements were in any case based on  $C^0$  shape functions – thus this development was welcome in that universally valid  $C^0$  shape functions and their derivatives could be used for a very wide range of structural applications.

However, surprisingly dramatic failures came to be noticed when  $C^0$  elements were formulated. The greater part of academic activity in the late seventies, most of the eighties and even in the nineties was spent in understanding and eliminating what were called the locking problems. A good idea of the challenges involved can be seen in two recent reviews – a bibliography of the finite element formulation of constrained media elasticity (Prathap & Nirmala 1990) – about 500 papers in thirty years, and a review of the quest for a reliable degenerate shell element (Gilewski & Radwanska 1991) – over 350 papers in about three decades of activity.

These spectacular failures were called the 'locking' problems in  $C^0$  finite elements. It was not clear why the displacement type method, as it was understood around 1977, should produce for such problems, answers that were only a fraction of a percent of the correct answer with a practical level of discretisation. Studies in recent years have established that an aspect known as consistency must be taken into account.

The consistency paradigm requires that the interpolation functions chosen to initiate the discretisation process must also ensure that any special constraints that are anticipated must be allowed for in a consistent way. Failure to do so causes solutions to lock to erroneous answers. The paradigm showed how elements can be designed to be free of these errors. It also enabled error-analysis procedures that allowed errors to be traced to the inconsistencies in the representation to be developed. The authors have now developed a family of such error-free robust elements for application in structural mechanics and these are now available in a package, FEPACS (finite element package for analysis of composite structures), developed at the National Aerospace Laboratories.

This article would therefore review the understanding of such errors on a paradigmatic basis and the arrival at the end, of a family of robust elements of acceptable accuracy for use in typical GPP.

## 2. Difficulties with $C^0$ elements: Locking and stress oscillations

With the wide-spread acceptance of the  $C^0$  family of elements, instances where the finite elements models of practical structures under certain physical conditions produced very erroneous solutions in spite of satisfying the continuity and completeness requirements came to be noticed (Doherty *et al* 1969; Pawsey & Clough 1971; Zienkiewicz *et al* 1971). Such errors are today classified as locking – where errors in solutions grow indefinitely as the physical limits are approached (Prathap & Bhashyam 1982)

and delayed convergence – where the convergence rate of the solutions is much lower than that assured by the conventional continuity and completeness requirements of the finite element method (Prathap & Babu 1986b). Such errors in displacement solution are always associated with violent stress oscillations (Prathap & Babu 1987).

Walz *et al* (1970) classified errors into two categories – errors of first kind (errors due to the discretisation process but which disappear rapidly as mesh is improved) and errors of the second kind (discretisation errors which disappear very slowly and which get exaggerated when some structural parameter is changed). At the time this classification appeared (Walz *et al* 1970) it was not known that the latter class of errors were due to incorrect representation of the constrained strain energy components and arise purely from the way the finite element fields are expressed. Recent work shows that such errors become very serious in a particular class of problem – constrained media elasticity (Babu 1985; Prathap 1986, 1993; Naganarayana 1991). These problems span a wide range of structural phenomena – shear-flexible beams using Timoshenko theory (Prathap & Bhashyam 1982) and plates/shells (Mindlin theory) suffering from shear locking (Hughes *et al* 1977; Prathap & Viswanath 1983; Bathe & Dvorkin 1985; Hinton & Huang 1986; Donea & Lamain 1987; Prathap & Somashekar 1988), curved beam/shell structures suffering from membrane locking (Stolarski & Belytschko 1981; Prathap 1985a, b; Prathap & Babu 1986a; Jang & Pinsky 1988), 2-D plane-stress, plane-strain and 3-D elasticity suffering from parasitic shear (Cook 1975; Prathap 1985c; Prathap *et al* 1986) and/or near incompressibility locking (Chandra & Prathap 1989; Naganarayana & Prathap 1991) etc. These are described in detail in table 1.

Many *ad hoc* techniques have been suggested to overcome such difficulties. Reduced/

**Table 1.** Some constrained-multi-strain problems in structural mechanics.

Class of structural problem	Strain fields		Type of constraints and the associated penalty limits
	Unconstrained	Constrained	
Plane stress Plane strain 3-D elasticity in modes of flexure	Normal $\epsilon$	Shear $\gamma$	$\gamma \rightarrow 0$ as $(b/l) \rightarrow 0$ where $b$ and $l$ form the section on which $\gamma$ is acting s.t. $b < l$
Plane stress Plane strain 3-D elasticity in modes of near incompressibility	Distortional $\epsilon_d$	Dilatational $\epsilon_v$	$\epsilon_v \rightarrow 0$ as $\mu \rightarrow 0.5$ for isotropic materials where $\mu$ is Poisson's ratio
Shear-flexible beam (Timoshenko) plates (Mindlin)	Bending $\chi$	Transverse shear $\gamma$	$\gamma \rightarrow 0$ as $(t/l) \rightarrow 0$ where $t$ = thickness and $l$ = "element length"
Curved beams and shells	Bending $\chi$	Membrane $\epsilon$	$\epsilon \rightarrow 0$ as $(Rt/l^2) \rightarrow 0$ where $t$ = thickness and $R$ = radius of curvature $l$ = "element length"

selective integration (Pawsey & Clough 1971; Zienkiewicz *et al* 1971; Zienkiewicz & Hinton 1976; Hughes *et al* 1977, 1978), assumed strain methods (MacNeal 1982; Oleson 1983), addition of bubble modes (Wilson *et al* 1973), residual energy balancing (Fried 1974, 1975), spurious mode decomposition (Belytschko *et al* 1984, 1985), discontinuous force-field mixed methods (Noor & Hartley 1977; Noor & Anderson 1982; Noor & Peters 1981), symbolic Fourier synthesis (Park 1984), unequal field interpolations with condensation of constraints (Tessler & Dong 1981), Kirchhoff mode method (Stolarski *et al* 1985), quasi-conforming techniques (Tang *et al* 1984), use of trigonometric basis functions (Hepper & Hansen 1987), using shear constraints (Crisfield 1984) etc. represent a broad coverage of the various artifices used with varying degrees of success to resolve these issues. Often these procedures lacked an explanation for their success. Sometimes they were successful in one context and failed in some other. Again the reason for such behaviour was not clear.

Most explanations for the locking behaviour available at the time we started our work – singularity of shear stiffness (Zienkiewicz 1977), constraint counting and rank of the shear stiffness matrix (Cook *et al* 1981; Hughes 1987) – lacked a rigorous scientific basis. Zienkiewicz (1977) argued that the elements which lock have non-singular shear stiffness matrices while the shear stiffness matrices of elements after reduced integration are singular. It was argued therefore that locking is due to non-singularity of the shear stiffness matrix and a reduced integration order that induces singularity in the shear stiffness matrix is recommended. In other words, the locking behaviour of an element is due to the *high rank* of the penalty-linked stiffness matrices. But, it was soon realised that an arbitrary reduction of the rank of the penalty-linked stiffness matrices may lead to the undesirable spurious zero energy mechanisms (Hughes 1987). The optimal rank for the shear stiffness matrix is often determined by a technique known as constraint counting (Malkus & Hughes 1978). The method of constraint counting makes an attempt to determine optimal integration order based on number of constraints given in a problem. It is based on the ratio  $r$  of the total active degrees of freedom in a given mesh  $n$  to the total number of penalty constraints  $m$  ( $r = n/m$ ). Locking occurs if  $r \leq 1$ . The mesh does not lock if  $r$  is *slightly* greater than unity and may have spurious zero energy modes if it is *too high*. Heuristically it argued that the near-optimal ratios are  $r = 2/1$  for 2-D problems and  $r = 3/1$  for 3-D problems (Cook *et al* 1981).

It should be observed here that such explanations are heuristic and lack a rigorous scientific basis i.e. the validity of the explanations is not numerically verifiable (falsifiable) since a causal relationship between the locking errors and the rank of the 'constrained' stiffness matrix is not established. Such arguments often follow from the given mesh for a structural problem rather than the basis of discretisation adopted in the finite element formulation. Such an explanation, hence, cannot be generally applied in developing a finite element. These arguments cannot identify the milder problems of delayed convergence which are observed in the case of higher order elements. Finally, they attempt explanations based on the symptoms accompanying the problem of locking (i.e. the high rank of non-singularity of the shear stiffness matrix is a symptom of an inconsistent formulation) rather than exploring the cause for the errors. There is no established procedure for obtaining the optimal rank of the penalty-linked stiffness matrices in literature. Here, we review some of the work done to provide a scientific basis for the origin of such errors. Very recently, it was possible to relate the consistency paradigm and the requirements that follow from it to the rank of the penalty-linked stiffness matrix, showing that there is a link

between the cause of such errors and the symptoms (i.e. high rank, non-singularity of matrices etc.) associated with the locking problem (Prathap 1994).

In the next sections, we shall present a scheme for error-free displacement type finite element formulation based on the consistency and correctness paradigms. The elements developed thus are now incorporated in an in-house finite element package called FEPACS.

### 3. Field-consistency

Prathap & Bhashyam (1982) demonstrated that it is the *inconsistent* finite element representation of the constrained state of the strain energy in the respective penalty limits that causes problems like locking and stress oscillations. It considered a Timoshenko beam element formulation and showed that the constraint of vanishing transverse shear strain energy near vanishing beam thickness imposed two types of constraints. The constraints which had contributions from all the displacement fields appearing in the respective strain field definitions were classified as true constraints and the constraints that do not have contributions from at least one of the constituent displacement fields were classified as spurious constraints. It was also demonstrated that the latter (i.e. spurious constraints) disturbed the bending strain energy in the penalty limits. Using this fact, analytical *a priori* error estimates were constructed to prove that it was the spurious constraints that caused locking and the associated stress oscillations (Prathap & Babu 1986b, 1987). This was confirmed by conducting appropriate numerical experiments with the *inconsistent* finite element formulations which contain spurious constraints.

Eventually, the field-consistency paradigm emerged (Prathap 1986) to provide an explanation as well as a remedy for difficulties like locking and stress oscillations in the finite element formulation of the problems in constrained media elasticity. It holds the inconsistent representation of the constrained state of the corresponding strain energy components, giving rise to the spurious constraints in the penalty limits, responsible for such difficulties. It also suggests that the energy components associated with the spurious constraints have to be eliminated from the formulation for removal of the above-mentioned difficulties from the element formulation. The paradigm can be broadly stated as follows.

In a constrained media problem, some strains will have to vanish under certain conditions. Strain fields derived from displacement shape functions cannot always do this in a meaningful manner – spurious constraints are generated which cause locking. The consistency condition demands that the discretised strain field interpolations must be so constituted that it will enforce only physically true constraints when the discretised functionals for the strain energy of a finite element are constrained.

In the development of a finite element, the field variables are interpolated using interpolations of a certain order. From these definitions, one can compute the strain fields using the strain-displacement relations. These are obtained as interpolations associated with the constants that were introduced in the field variable interpolations. Depending on the order of the derivatives of each field variable appearing in the definition of that strain field (e.g. the shear strain in a Timoshenko theory will have  $\theta$  and the first derivative of  $w$ ), the coefficients of the strain field interpolations may have constants from all the contributing field variable interpolations or from only one or some of these. In some limiting cases of physical behaviour, these strain fields

can be constrained to be zero values, e.g. the vanishing shear strain in a thin Timoshenko beam. Where the strain-field is such that all the terms in it (i.e. constant, linear, quadratic etc.) have, associated with it, coefficients from all the independent interpolations of the field variables that appear in the definition of that strain-field, the constraint that appears in the limit can be correctly enforced. We shall call such a representation *field-consistent*. The constraints thus enforced are *true constraints*. Where the strain-field has coefficients in which the contributions from some of the field variables are absent, the constraints may incorrectly constrain some of these terms. This *field-inconsistent* formulation is said to enforce additional *spurious constraints*.

We shall also determine procedures that can modify the element characteristics so that the consistency requirements are met. We shall call such elements the *field-consistent* elements as opposed to the *field-inconsistent* elements which do not take into account such requirements. There is a unique manner in which the field-consistent elements have to be generated – they have to satisfy a condition we shall call the *correctness* condition which will ensure that the variational theorems are not violated in the process of modifying the element stiffness matrix.

Scientific evidence and analytical proof for this paradigm have been provided and verified through several practical examples of finite element formulations over the years (Babu 1985; Prathap 1986, 1993; Naganarayana 1991). Using the so-called field-reconstitution technique, analytical *a priori* estimates of the errors in displacement as well as stress recovery arising due to violation of various field-consistency requirements are derived for a family of elements and are digitally verified using appropriate computational models, see Prathap & Nirmala (1990) for a bibliography.

Most of the *ad hoc* techniques mentioned before offer procedures for achieving field-consistency with varied degrees of success. The assumed strain methods appear to be the most versatile among these since they have the capability of isolating the spurious constraints in the formulation. Thus the behaviour of the inconsistent terms in the constrained strain fields and the associated spurious constraints becomes apparent and the construction of *a priori* error estimates, with reference to displacement as well as stress recovery, becomes simplified. The assumed strain methods essentially try to eliminate the inconsistent terms in the original strain field (derived by the gradient operations on the kinematically admissible displacement fields) for the constrained strain component. The popular procedures used in the literature for achieving this are reduced/selective integration (Zienkiewicz & Hinton 1976; Hughes *et al* 1978), least squares method (Bose & Kirkhop 1984), mean value method (Donea & Lamain 1987), collocation of the strain fields at certain standard points (Huang & Hinton 1984; Bathe & Dvorkin 1985) etc. and all these belong to assumed strain methods in a broad sense of definition.

Various aspects of the field-consistency paradigm are briefly illustrated using the simple example of a 2-noded  $C^0$ -continuous Timoshenko beam element. The field-reconstitution technique is used to derive *a priori* analytical error estimates for the additional stiffness parameter and stress oscillations which can be digitally verified. The technique is used to derive *a priori* error norms for many other useful elements in a series of publications.

### 3.1 Example—linear Timoshenko beam element (BEAM2)

A 2-noded beam element with two degrees of freedom (deflection  $w$  and section rotation  $\theta$ ) per node, shown in figure 1, is considered.

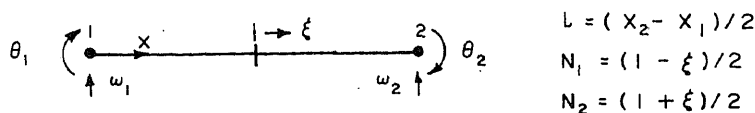


Figure 1. 2-noded beam element.

Using the linear shape functions  $N_1$  and  $N_2$  the displacement fields can be expressed as,

$$\theta = a_0 + a_1 \xi \quad \text{and} \quad w = b_0 + b_1 \xi \quad (1)$$

Now, the two strain fields, flexure  $\chi$  and the transverse shear  $\gamma$ , can be derived using (1) as

$$\begin{Bmatrix} \chi \\ \gamma \end{Bmatrix} = \begin{Bmatrix} d\theta/dx \\ \theta - dw/dx \end{Bmatrix} = \begin{Bmatrix} a_1/l \\ (a_0 - b_1/l) + a_1 \xi \end{Bmatrix}. \quad (2)$$

The strain energy can be considered as the sum of bending and shear strain energy components as follows.

$$\begin{aligned} U &= U_B + U_S = \int (El/2)(\chi)^2 dx + \int (kGA/2)(\gamma)^2 dx, \\ &= (El)(a_1/l)^2 + (kGA) \{ (a_0 - b_1/l)^2 + (a_1)^2/3 \}. \end{aligned} \quad (3)$$

In the penalty limits of vanishing beam thickness, the shear strain energy  $U_S$  vanishes resulting in two constraints,

$$a_0 = b_1/l \equiv \theta = dw/dx, \quad (4)$$

$$a_1 = 0 \equiv d\theta/dx = 0. \quad (5)$$

It is apparent that (4) is the true constraint representing the Euler-Bernoulli condition in the thin limits. But (5) is the spurious constraint which directly disturbs the bending strain energy. It can be noted that the spurious constraint corresponds to a situation where the beam is locked against rotation, thus explaining the locking behaviour of the problem. The field-consistency paradigm now suggests that the linear term in the shear strain field be eliminated for removal of locking and violent stress oscillations. The method of Legendre polynomial expansion can be used here so that the inconsistent linear term in shear strain is to be simply dropped. The procedure and variational justification for it are discussed in detail, in the next section.

It is also possible to obtain *a priori* error estimates by considering the strain energy after discretisation by a procedure called the functional reconstitution technique. For this example, it is possible to show that the effect of retaining the spurious constraint leads to an artificial stiffening of the bending action by a factor  $(kGA l^2/3EI)$ ; i.e. the discretised beam behaves as if the moment of inertia of the cross section has changed from  $I$  to  $I'$  such that,

$$I'/I = 1 + kGA l^2/3EI. \quad (6)$$

Thus an analytical error norm for the additional stiffness parameter  $e_{th}$  can be written

as,

$$e_{th} = I'/I - 1 = kGA l^2/3EI. \quad (7)$$

This can be compared with results obtained from a actual finite element computation as,

$$e_{fem} = \omega(th)/\omega(fem) - 1 \quad (8)$$

The agreement of the computed errors (8) with the predicted errors (7) is very good and has been documented in Prathap & Bhashyam (1982) and Babu (1985).

Through a similar exercise, in an inconsistent element with severe shear locking (when  $e \gg 1$ ), the shear force resultant can be expressed as,

$$Q = \bar{Q} + (3M_0/l)\zeta,$$

where  $\bar{Q}$  represents the correct shear force contributed from the consistent part of the shear strain field and  $M_0$  is the constant part of the exact bending moment distribution. Thus the shear force recovered from the inconsistent element will have violent linear oscillations which are proportional to the constant part of the bending moment  $M_0$ . Again, the analytical predictions of the stress oscillations are well-matched by the finite element results (Babu 1985).

Such behaviour, i.e. locking and stress oscillations, is typical of inconsistently formulated beam, plate/shell, plane stress/plane strain and brick elements, the work horse elements of all current general purpose packages. It is imperative therefore that the design of a family of such elements must be carefully done to ensure that during formulation, the inconsistencies are systematically identified and removed so that the elements are free of such errors. The field-consistency paradigm offers us a conceptual scheme which enables this to be done.

#### 4. Variational correctness

We now turn our attention to the task of formulating a consistent element without violating any of the norms required from energy and variational principles. It has become clear to us from the previous section that the field-consistency paradigm suggests the form of the interpolation functions for the constrained strain fields. However, the reconstituted *assumed* strain field interpolations cannot be chosen arbitrarily (as done, for example, in Mohr 1982). A variational basis for the *correct* field-redistribution (that is to find the coefficients of the consistent field from those of the original field) can be obtained by determining the conditions for exact equivalence of the assumed strain displacement approach to the mixed approaches based on the Hellinger-Reissner or Hu-Washizu variational theorems (Simo & Hughes 1986; Prathap 1988). The coefficients of the *field-consistent* assumed strain fields should now be determined from an orthogonality condition that arises when the equivalence of the minimum total potential energy principle with respect to the mixed variational principles is sought. This leads to another fundamental requirement in certain finite element formulations involving redistributed (or assumed) fields (e.g. strain or stress) – *variational correctness* – that the redistributed field should be orthogonal to the error introduced because of the field-redistribution (Prathap 1988; Prathap & Naganarayana



1988; Naganarayana & Prathap 1989a):

$$\int \delta \bar{\gamma} (\bar{\gamma} - \gamma) dx = 0,$$

where  $\gamma$  represents the constrained strain field kinematically derived from the constituent displacement fields while  $\bar{\gamma}$  is the assumed field to be determined from the original field such that  $\bar{\gamma}$  is field-consistent and the resulting element formulation is free of locking and stress oscillations.

If the orthogonality condition that represents this equivalence is violated while designing the strain field to satisfy the field-consistency criteria alone, the resulting elements have poorer efficiency and are also plagued by undesirable strain and stress oscillations (Prathap & Naganarayana 1988; Naganarayana & Prathap 1989a).

In Prathap & Naganarayana (1988), the variationally correct (or orthogonal) and incorrect (or non-orthogonal) field-consistent assumed strain forms of the quadratic and cubic shear deformable beam elements are used to explore these aspects in detail. It is shown that the non-orthogonal formulations can lead to reasonably accurate displacement solutions but have spurious stress oscillations. Using the field-reconstitution technique, *a priori* analytical estimates were derived for the magnitude and pattern of these stress oscillations and tested digitally using computational results in Prathap & Naganarayana (1988). These stress oscillations are related to the presence of artificially created spurious load mechanisms which are self-equilibrating.

These spurious load mechanisms lead to additional spurious linear oscillations in the shear force and bending moment in a quadratic Timoshenko beam element and spurious quadratic oscillations in bending moment (but no change in the shear forces) in cubic Timoshenko beam element as demonstrated in Prathap & Naganarayana (1988). The resulting extraneous oscillations in the stress fields may often lead to difficulty in identifying points for optimal stress recovery in the element domain e.g. line-consistent 8-noded plate element (Naganarayana & Prathap 1989a).

In many problems, it is possible to determine the variationally correct strain field by expanding the inconsistent field in terms of Legendre polynomials; it is easy then to identify and eliminate the inconsistent terms. The logic is simple and direct; the coefficients associated with each Legendre polynomial represents a discretised constraint in the penalty limits; identify the inconsistent Legendre term applying the field-consistency paradigm and simply drop it to get the correct and consistent field. Since the Legendre polynomials are orthogonal in the domain of integration, it follows that this procedure satisfies the orthogonality condition arising from the equivalence of the minimum total potential energy principles with reference to the mixed principles and hence the method is variationally correct. The same result can often be achieved quite simply by using reduced integration if  $\gamma$  is *one order* higher than  $\bar{\gamma}$ . It will be variationally incorrect otherwise. Reduced integration is universally popular since it is very easy to implement it on a computer. We shall see later that reduced integration cannot satisfy the edge-consistency requirements and hence fails when used in a distorted mesh. Hence the methods like Legendre polynomial expansion and truncation become very important in development of general purpose finite elements.

## 5. Consistent non-uniform mapping and edge-consistency

So far, we have looked at the consistency requirements in a simplified form that can be directly applied only to straight line elements or elements of rectangular form. It

is not practical to have such a restriction for the plate/shell elements in a general purpose application. The development of an efficient and robust quadrilateral plate bending element has therefore been a formidable challenge. The point here is to ensure that consistency is maintained in curved and arbitrary quadrilateral forms of an element. The current practice in the development of elements is to use what is called the isoparametric concept – this requires the use of a covariant or natural co-ordinate system for all interpolations – i.e. the same interpolations serving to map both geometry (from the natural system to the Cartesian system) and displacements. The strains are based on the Cartesian system but interpolated using natural co-ordinates. It is therefore necessary to allow for changes from one system to another without loss of consistency, especially as far as the mapping of the constrained strain fields are concerned. It turned out that one crucial factor was the way tangential strain components on each edge of an element had to be defined under situations of non-uniform mapping.

That small errors in data preparation leading to distortion of the element can cause large errors in the solutions was reported by Hoppe (Hoppe 1984, 1985). This is due to the non-uniform mapping from the covariant system to the Cartesian system. The effect in the case of a constrained media problem is much more severe. Since curved beam, and quadrilateral plate and shell elements will intrinsically have such mapping conditions, there can be very large errors on this account. It is therefore necessary to take care to design an element so that accuracy and efficiency when used in a distorted mesh do not go down rapidly. There have been several attempts in literature to develop elements that are free of locking in their general form (Hughes *et al* 1977; MacNeal 1982; Hinton & Huang 1986; Donea & Lamain 1987) – most of them using the sampling points on the element edges. However, they lack an explanation for their success.

It was observed in Prathap & Naganarayana (1992) that it is not sufficient if field-consistency is achieved in a variationally correct sense in the covariant natural system alone in such cases. It was demonstrated that the difficulties are with the non-uniform mapping of the strain fields because of which the consistency and correctness conditions achieved in the natural covariant system are not preserved over the element domain and across its boundary after transformation into the Cartesian system. Several methods of achieving *consistent mapping* of the constrained strain fields from the natural system to the Cartesian system were discussed from two popular points of view – Cartesian base formulation and covariant base formulation.

The additional requirement for an element to be free of errors in a *patch* is the *edge-consistency* requirement (Prathap & Somashekar 1988). It was shown that the *tangential* strain components which are continuous across the element boundary in the undistorted covariant natural system should remain continuous even after the necessary transformations, and that the *tangential* strain components should be built from their corresponding *tangential* displacement components only. Though the Cartesian base formulation appeared to be very accurate in Prathap & Naganarayana (1992), identification of *tangential strain* components and hence achieving edge-consistency becomes a formidable job. Thus a covariant base formulation becomes desirable for developing general purpose finite elements.

We should note here that, in case of covariant base formulations, consistent mapping preserves the field-consistency over the element domain in the Cartesian system. However, special methods have to be used if edge-consistency has to be satisfied. An effective (from both accuracy and computational points of view) procedure namely, *nodal Jacobian transformation*, is developed and employed to

achieve both consistent mapping and edge-consistency while developing several general purpose finite elements (Prathap & Somashekar 1988; Prathap *et al* 1988; Naganarayana & Prathap 1989a, 1989b; Naganarayana *et al* 1992; Prathap & Naganarayana 1992). Such elements which satisfy both field- and edge-consistency requirements are subjected to several severe patch tests and are found to be very accurate.

It is interesting to observe here that the methods of sampling the constrained *tangential* strain components at the error-free points on the element boundary (Hughes *et al* 1977; MacNeal 1982; Hinton & Huang 1986; Donea & Lamain 1987) in fact, try to achieve field- and edge-consistency requirements in the Cartesian system in a similar fashion. This explains their success when used in a distorted mesh.

## 6. Problems with initial strain/stress and varying moduli

In problems with initial stress/strain or with varying moduli, the discretised stress-resultant and strain fields will be of different order – consequently there is a loss of consistency in the formulation that results in oscillations in the stress-resultant fields (Prathap & Naganarayana 1990a, and to be published; Naganarayana 1991). For example, it has been known for some time that thermal stresses computed directly from stress-strain and strain-displacement matrices in a finite element analysis with simple finite elements can show large errors (Ojalvo 1974; Pitt & Hartl 1980). The conventional wisdom to tackle such problems is to sample stresses/strains at the Gauss points corresponding to a reduced integration rule. However, cases exist, e.g. the tapered quadratic bar element (Prathap & Naganarayana 1990a), where the oscillations are such that easily identifiable points of accurate stress do not exist.

In the computation of the stiffness matrix, energy terms of the form  $U = \int \sigma^T \epsilon d\Omega$  exist and if the order of  $\sigma$  ( $\sigma$  may be stress or stress-resultant depending on the problem) is higher than that of the strain  $\epsilon$ , the higher order term may not “do work” on the strain terms and are not recognised in the stiffness matrix. Thus, stress-resultants computed from the displacements recovered from such a formulation show extraneous oscillations.

It is necessary to define a consistent stress or stress-resultant field to ensure that the stress recovery reflects correctly the order of strain interpolations used. The orthogonality conditions, required for reconstituting the assumed fields for both strain and stress functions (from their original fields) simultaneously, can be obtained from the Hu-Washizu variational principle as,

$$\int \delta \bar{Q} (\bar{\gamma} - \gamma) d\Omega = 0; \quad \int \delta \bar{\gamma} (D\bar{\gamma} - \bar{Q}) d\Omega = 0,$$

where,  $\gamma, \bar{\gamma}, D\bar{\gamma}$  and  $\bar{Q}$  represent the original inconsistent strain field, consistent strain field, stress-resultant field derived from the consistent strain field through constitutive laws and the consistent stress resultant field respectively.

In a varying moduli problem, due to varying  $D$ , the strain and stress-resultant fields will be of different interpolation order. In a thermal stress problem, the initial strain field in the element  $\epsilon_0$  which will be of the same order as the temperature field in the element, and the total strain field, which is obtained by differentiating the displacement fields will be of different order, especially if the temperature fields vary

significantly over the domain and are interpolated by the same isoparametric functions as the displacement fields. Consequently, the total stress will be of an order higher than the strain field. Therefore, in recovering stresses in such problems, care must be taken to maintain consistency. In our work, the definitions of stress-resultants, or initial strains/stresses are *a priori* made consistent with the strain fields by invoking the orthogonality conditions seen earlier in this section.

## 7. A complete consistent and correct procedure for displacement type finite element formulation

So far, we have reviewed the important consistency and correctness paradigms necessary for finite element formulation free of locking and stress oscillations. Here, we sum up all these concepts and the additional considerations that are necessary for developing a general-purpose finite-element library for analysis of composite structures. For continuity, we recapitulate our understanding of the sources for the locking and stress oscillations and the required consistency and correctness paradigms to eliminate the same (table 2).

*Continuity:* The displacement function must be continuous over the element domain and across the element boundary in a given arbitrary finite element mesh.

**Table 2.** Types of errors, their sources in finite element analysis and the associated paradigms.

Source	Symptoms	Paradigm/concepts
Finite element discretisation	<i>Errors of first kind</i> Discretisation errors (errors of first kind)	Continuity and Completeness
Constrained media elasticity	<i>Errors of second kind</i> Locking, delayed convergence, stress oscillations	Field-consistency
Element distortion (nonuniform mapping)	Locking, delayed convergence, stress oscillations	Edge-consistency and consistent mapping
Varying moduli	Stress oscillations	Stress field-consistency
Initial strain & stress field representation	Strain and stress oscillations	Stress field-consistency
Redistribution of strain/stress fields	Poorer convergence spurious load mechanisms and stress oscillations	Variational correctness
Modelling warped Surface with linear elements	Erroneous displacements and stress recovery	Warping correction

achieve both consistent mapping and edge-consistency while developing several general purpose finite elements (Prathap & Somashekar 1988; Prathap *et al* 1988; Naganarayana & Prathap 1989a, 1989b; Naganarayana *et al* 1992; Prathap & Naganarayana 1992). Such elements which satisfy both field- and edge-consistency requirements are subjected to several severe patch tests and are found to be very accurate.

It is interesting to observe here that the methods of sampling the constrained *tangential* strain components at the error-free points on the element boundary (Hughes *et al* 1977; MacNeal 1982; Hinton & Huang 1986; Donea & Lamain 1987) in fact, try to achieve field- and edge-consistency requirements in the Cartesian system in a similar fashion. This explains their success when used in a distorted mesh.

## 6. Problems with initial strain/stress and varying moduli

In problems with initial stress/strain or with varying moduli, the discretised stress-resultant and strain fields will be of different order – consequently there is a loss of consistency in the formulation that results in oscillations in the stress-resultant fields (Prathap & Naganarayana 1990a, and to be published; Naganarayana 1991). For example, it has been known for some time that thermal stresses computed directly from stress-strain and strain-displacement matrices in a finite element analysis with simple finite elements can show large errors (Ojalvo 1974; Pittr & Hartl 1980). The conventional wisdom to tackle such problems is to sample stresses/strains at the Gauss points corresponding to a reduced integration rule. However, cases exist, e.g. the tapered quadratic bar element (Prathap & Naganarayana 1990a), where the oscillations are such that easily identifiable points of accurate stress do not exist.

In the computation of the stiffness matrix, energy terms of the form  $U = \int \sigma^T \epsilon d\Omega$  exist and if the order of  $\sigma$  ( $\sigma$  may be stress or stress-resultant depending on the problem) is higher than that of the strain  $\epsilon$ , the higher order term may not “do work” on the strain terms and are not recognised in the stiffness matrix. Thus, stress-resultants computed from the displacements recovered from such a formulation show extraneous oscillations.

It is necessary to define a consistent stress or stress-resultant field to ensure that the stress recovery reflects correctly the order of strain interpolations used. The orthogonality conditions, required for reconstituting the assumed fields for both strain and stress functions (from their original fields) simultaneously, can be obtained from the Hu-Washizu variational principle as,

$$\int \delta \bar{Q} (\bar{\gamma} - \gamma) d\Omega = 0; \quad \int \delta \bar{\gamma} (D\bar{\gamma} - \bar{Q}) d\Omega = 0,$$

where,  $\gamma$ ,  $\bar{\gamma}$ ,  $D\bar{\gamma}$  and  $\bar{Q}$  represent the original inconsistent strain field, consistent strain field, stress-resultant field derived from the consistent strain field through constitutive laws and the consistent stress resultant field respectively.

In a varying moduli problem, due to varying  $D$ , the strain and stress-resultant fields will be of different interpolation order. In a thermal stress problem, the initial strain field in the element  $\epsilon_0$  which will be of the same order as the temperature field in the element, and the total strain field, which is obtained by differentiating the displacement fields will be of different order, especially if the temperature fields vary

significantly over the domain and are interpolated by the same isoparametric functions as the displacement fields. Consequently, the total stress will be of an order higher than the strain field. Therefore, in recovering stresses in such problems, care must be taken to maintain consistency. In our work, the definitions of stress-resultants, or initial strains/stresses are *a priori* made consistent with the strain fields by invoking the orthogonality conditions seen earlier in this section.

## 7. A complete consistent and correct procedure for displacement type finite element formulation

So far, we have reviewed the important consistency and correctness paradigms necessary for finite element formulation free of locking and stress oscillations. Here, we sum up all these concepts and the additional considerations that are necessary for developing a general-purpose finite-element library for analysis of composite structures. For continuity, we recapitulate our understanding of the sources for the locking and stress oscillations and the required consistency and correctness paradigms to eliminate the same (table 2).

**Continuity:** The displacement function must be continuous over the element domain and across the element boundary in a given arbitrary finite element mesh.

**Table 2.** Types of errors, their sources in finite element analysis and the associated paradigms.

Source	Symptoms	Paradigm/concepts
Finite element discretisation	<i>Errors of first kind</i> Discretisation errors (errors of first kind)	Continuity and Completeness
Constrained media elasticity	<i>Errors of second kind</i> Locking, delayed convergence, stress oscillations	Field-consistency
Element distortion (nonuniform mapping)	Locking, delayed convergence, stress oscillations	Edge-consistency and consistent mapping
Varying moduli	Stress oscillations	Stress field-consistency
Initial strain & stress field representation	Strain and stress oscillations	Stress field-consistency
Redistribution of strain/stress fields	Poorer convergence spurious load mechanisms and stress oscillations	Variational correctness
Modelling warped Surface with linear elements	Erroneous displacements and stress recovery	Warping correction

**Completeness:** The strain/stress fields should be able to model strain-free rigid body motion of the element and the constant strain state of the element deformation.

**Field-consistency:** The terms in a constrained strain field that have partial contribution from the constituent displacement fields, leading to the spurious constraints in the penalty limits for the corresponding strain energy components, should be eliminated from the formulation for assuring convergence of results from a finite element model of the structural problems belonging to the class of constrained media elasticity.

**Consistent mapping:** Mapping of the strain/stress fields from the covariant natural system (where the element configuration is always undistorted) to the Cartesian system should not introduce any additional spurious constraints.

**Edge-consistency:** The *tangential* strain components which are continuous across the element boundary in the undistorted covariant natural system would remain continuous even after the necessary transformations; and the *tangential* strain components should be built from their corresponding *tangential* displacement components only.

**Stress field-consistency:** The terms in the strain and/or stress fields that do not participate in the strain energy computations (and hence in the displacement recovery) should be eliminated while recovering the corresponding strain and/or stresses in a displacement type formulation.

**Variational correctness:** The redistributed field should be orthogonal to the error introduced because of the field-redistribution with reference to the original field.

The ideal characteristics of a finite element formulation for general purpose applications and the paradigms/concepts required to achieve the same in a scientific manner are briefly shown in table 3.

Based on the consistency and correctness principles several linear and quadratic elements, free of locking and stress oscillation, are developed for 1-, 2- and 3-dimensional applications.

## 8. FEPACS – finite element package for analysis of composite structures

FEPACS is a medium-sized general purpose package (i.e. about 20,000 lines of FORTRAN code) for the finite element analysis of isotropic, anisotropic and layered-composite structures. It has a family of simple, accurate and robust field-consistent elements. The package was initially built around the data and program organisation of SAP-IV, but recently, these have been modified and new solution capabilities are being introduced, to give FEPACS a character of its own.

The structural analysis program (SAP) is one of the earliest general purpose programs used for structural analysis through the finite element method. It was developed by Bathe *et al* (1974) under the sponsorship of many international organisations. The first version of SAP was released in September 1970 (Wilson 1970). The improved version, SAP-IV, which can be used for linear static and dynamic analysis of 3-dimensional structures was released in 1974 (Bathe *et al* 1974). As it had been released with its source code in the public domain, it has served as the spring board for many other finite element packages which were improved versions of the original package. It is well known that SAP-IV has reasonably efficient solution capabilities and data-handling procedures that can solve large 3-dimensional systems. However, its main weakness is its very old, outdated element library based on mainly linear finite elements for isotropic structures.

**Table 3.** Ideal characteristics of a finite element formulation.

Characteristics	Paradigms/concepts/procedures
Discretisation errors should <i>vanish</i> as the mesh is refined	Continuity and completeness
The element should be free of locking and/or delayed convergence	Field-consistency and Variational correctness
Element should be able to give variationally correct stress distribution without any oscillations	Field-consistency, Stress field-consistency, Variational correctness
Distorted geometry should not disturb the element geometry	Consistent mapping and edge-consistency
The element should be free of spurious zero energy mechanisms	Consistency and correctness Integration rules and Spectral analysis
Element formulation should be scientific & should provide <i>a priori</i> methods of error analysis (i.e. should not be based on numerically adjusted factors, heuristic arguments etc.)	Consistency and correctness paradigms supported by the functional reconstitution techniques
Element performance should be free from its geometry and position in space	Edge-consistency, appropriate local coordinate system and tensorial transformations

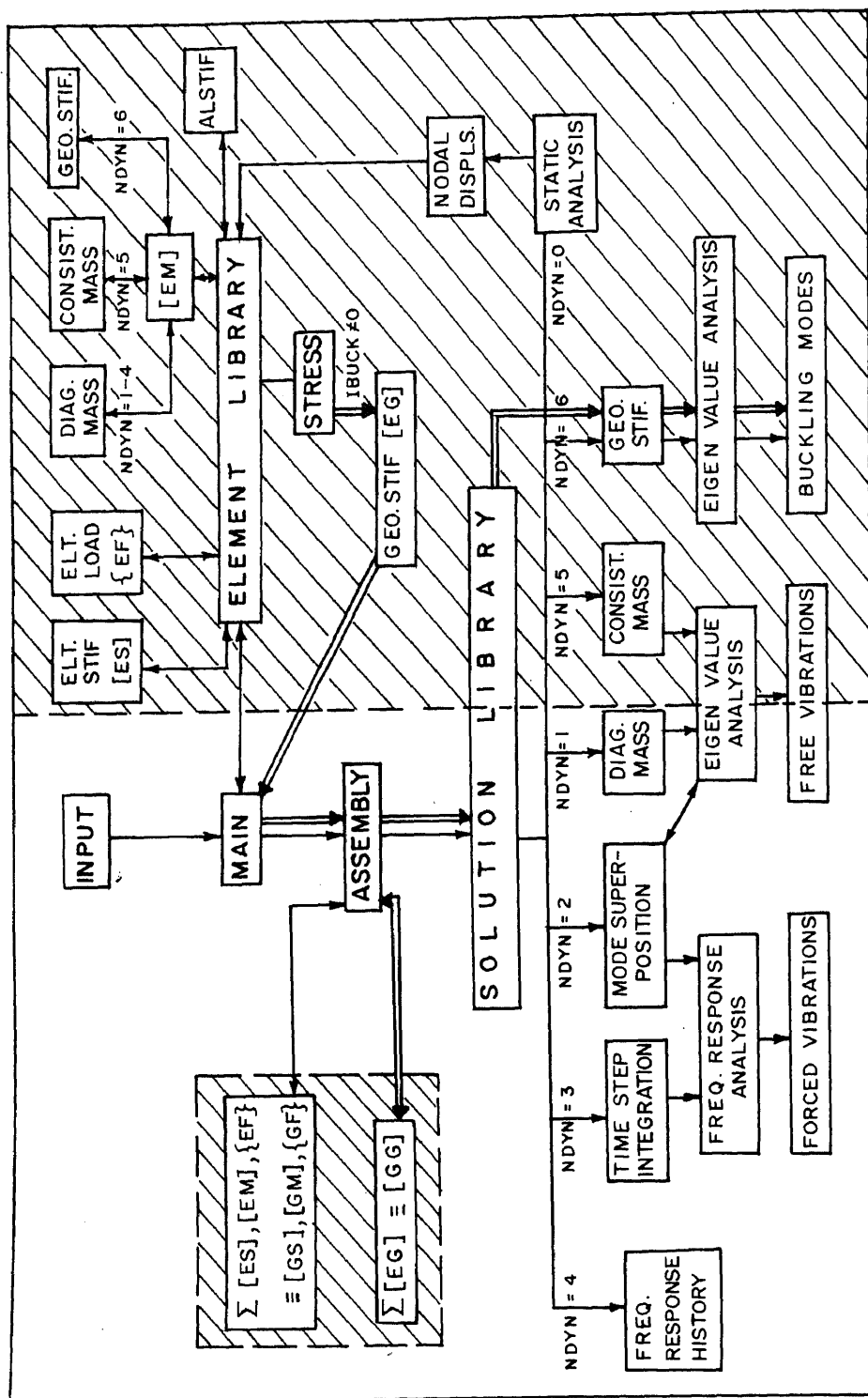
### 8.1 Finite element library

FEPACS has replaced this with the state-of-the-art field-consistent linear and quadratic 1-D, 2-D and 3-D element formulations – thus making the element library *complete* (Prathap *et al* 1989). The emphasis is on laminated anisotropic material structure (any other material constitution becomes a sub-set of this). A complete flow diagram representing the package is shown in figure 2.

Each element in FEPACS utilises field- and edge-consistent stiffness matrices derived in a variationally correct fashion; consistent initial (thermal) strain representation; consistent stress resultant field representation; diagonal or consistent mass matrices; and geometric stiffness matrices which can be used for linear buckling analysis; material constitutive laws are derived assuming general orthotropic layers. The finite elements included in the library are given below.

- (1) SPRING – 2-noded constant stiffness spring element,
- (2) TRUSS – 2-noded linear truss element,
- (3) BEAM2-T – 2-noded laminated linear Timoshenko beam element,
- (4) BEAM2-E – 2-noded laminated linear classical beam element,
- (5) BEAM3 – 3-noded laminated quadratic curved Timoshenko beam element (with taper and twist)
- (6) PLAXTQ – family of plane strain, plane stress & axisymmetric elements of triangular or quadrilateral shapes (orthotropic),
- (7) SHEL4 – 4-noded shear deformable laminated anisotropic Mindlin plate/ plane-shell element of quadrilateral shape with warping corrections,





**Figure 2.** Finite element package for analysis of composite structures (FEPACS).

- (8) SHEL8 – 8-noded quadrilateral laminated degenerated shell element,
- (9) HEXA8 – 8-noded linear solid brick element (orthotropic),
- (10) HEXA27 – 27-noded quadratic brick element (isotropic),

## 8.2 Library of solution capabilities

An eigenvalue solution routine based on the determinant search method has been added to FEPACS so that dynamic eigenvalue analysis using consistent mass matrices and buckling analysis of the linear structures can be carried out with the software. The solution routines can be efficiently used for both small-scale as well as large-scale problems using the respective in-core and out-of-core solution algorithms that are built in the program. The solution capabilities of FEPACS now include the following.

- (1) Static analysis under thermomechanical loads – Gauss elimination solution routines based on banded assembled matrices are used;
- (2) Natural frequency analysis – Both diagonal as well as consistent mass matrices can be used. Two eigenvalue solution routines, the determinant search method which is optimal for in-core solution and the subspace iteration method which is optimal for out-of-core solution, are available.
- (3) Natural frequency analysis followed by response history analysis – Only diagonal mass matrices need to be used. Method of superposition of the natural modes and the forced modes is utilised to get the structural response to dynamic loads.
- (4) Response history analysis using direct integration – Explicit time-step integration routines are used to get the structural response to dynamic loads without going through eigenvalue solution.
- (5) Natural frequency analysis followed by response spectrum analysis.
- (6) Buckling analysis – If the in-plane stresses are known *a priori* for each element, the analysis is done in a single run – construct the geometric stiffness matrices from the element in-plane stresses and solve for buckling loads and modes using the new eigenvalue solution routines. If they are not known (as in general structural analysis) the analysis is done in double runs – the in-plane stresses for each element are calculated from the first run of static analysis and the geometric stiffness matrices calculated from this static solution are then used for the second eigenvalue analysis.

## 8.3 Other features

Some of the other features specially incorporated in FEPACS to enhance its general purpose nature and its accessibility to users – for both regular analysis as well as for enhancement of the package with other capabilities include the following.

- (a) Generalised data input and output structures so that integration of pre- and post-processors can be efficiently achieved.
- (b) The program is organised in a highly modular fashion so that any new element or solution capability can be included with least difficulty.
- (c) The whole program is modified to use only *general* FORTRAN-77 commands such that the package is highly portable except for the scratch file operations, while are kept in a separate module which is compiler-dependent. The package is currently available on a variety of platforms like PC 386/486, workstations, super-mini

computers etc. with UNIX operating system, minimum 2 MB RAM and a FORTRAN compiler supporting scratch files of unlimited record length (only the scratch file allocations have to be modified). It can be loaded in any other operating system with similar support.

## 9. Conclusions

In this paper, many difficulties arising in displacement-type finite element formulation and the remedial measures offered for these have been reviewed. A general, complete and scientifically based procedure for formulating robust finite elements is provided. A general purpose finite element package (FEPACS) using a library of robust elements, developed along the lines presented here, is also reviewed. Though the emphasis is on the displacement-type finite formulations for structural analysis, the concepts discussed here are equally applicable to other types of finite element formulations applied to many other fields of engineering.

The authors are deeply indebted to Prof. R Narasimha, FRS, former Director of National Aerospace Laboratories, for his constant encouragement. They are also grateful to many colleagues – Dr H R Srinatha of Aeronautical Development Agency, Bangalore; Dr D H Bonde of Indian Space Research Organization, Bangalore and from the National Aerospace Laboratories, Drs G R Bhashyam, S Viswanath, S Chandra, the late Mr C Ramesh Babu, M/s B Sudhakar, K Guruprasad, S Nagaraj, V Baskar, Shaik Cheman, J Durga Prasad and Ms B R Shashirekha who have supported or worked on the development of FEPACS at various times.

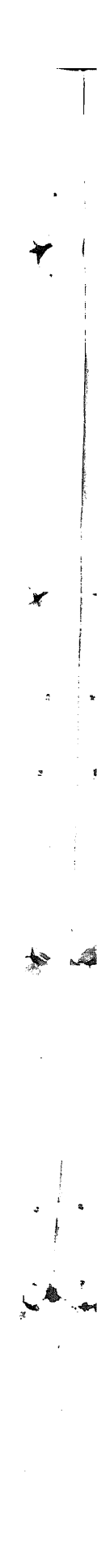
## References

- Babu C R 1985 *Field-consistency in the finite element formulation of multi-strain-field problems in structural mechanics*. Ph D Thesis, IIT Madras
- Bathe K J, Dvorkin E L 1985 A four-node plate bending element based on Mindlin/Reissner plate theory and a mixed interpolation. *Int. J. Numer. Methods Eng.* 21: 367–383
- Bathe K J, Wilson E L, Peterson F E 1974 SAP IV – A structural analysis program for static and dynamic response of linear systems. College of Engineering, University of California, Berkeley
- Belytschko T, Liu W K, Ong J S J 1984a Nine node Lagrange shell elements with spurious mode control. *Proc. AIAA/ASME. Struct. Dyn. Conf.*, Palm Springs
- Belytschko T, Liu W K, Ong J S J, Lamb D 1985 Implementation and application of a nine-noded Lagrangian shell element with spurious mode control. *Comput. Struct.* 20: 121–128
- Belytschko T, Ong J S J, Liu W K 1984b A consistent control of spurious singular modes in the 9-node Lagrangian element for the Laplace and Mindlin plate equations. *Comput. Methods Appl. Mech. Eng.* 44: 269–295
- Bose C J, Kirkhop J 1984 Least squares strain smoothing for the eight-node serendipity plane stress element. *Int. J. Numer. Methods Eng.* 20: 1164–1166
- Chandra S, Prathap G 1989 A field-consistent formulation for the 8-noded solid finite element. *Comput. Struct.* 33: 345–355
- Cook R D 1975 Avoidance of parasitic shear in plane element. *J. Struct. Div. ASCE*. 93: 43–66

- Cook R D, Malkus D S, Plesha M E 1981 *Concepts and applications of finite element analysis* 3rd edn (New York: John Wiley & Sons)
- Crisfield M A 1984 A quadratic Mindlin element using shear constraints. *Comput. Struct.* 18: 833–852
- Doherty W P, Wilson E L, Taylor R L 1969 Stress analysis of axisymmetric solids using higher order quadrilateral elements, SESM Report 69-3, Dept. of Civil Eng., University of California, Berkeley
- Donea J, Lamain L G 1987 A modified representation of transverse shear  $C^0$  quadrilateral plate elements. *Comput. Methods Appl. Mech. Eng.* 63: 183–207
- Fried I 1974 Finite element analysis of incompressible material by residual energy balancing. *Int. J. Solids Struct.* 10: 993–1002
- Fried I 1975 Finite element analysis of thin elastic shells with residual energy balancing and the role of the rigid body modes *J. Appl. Mech.* 42: 99–104
- Gilewski W, Radwanska M 1991 A survey of finite element models for the analysis of moderately thick shells. *Finite Element Anal. Design* 9: 1–21
- Heppler G R, Hansen J S 1987 Timoshenko beam finite elements using trigonometric basis functions. *AIAA J.* 26: 1378–1386
- Hinton E, Huang H C 1986 A family of quadrilateral Mindlin plate elements with substitute shear strain fields. *Comput. Struct.* 23: 409–431
- Hoppe V 1984 Errors in the finite element solutions introduced by mapping in isoparametric elements. *Proc. of the 4th world congress and exhibition on finite element methods* (Wimborne, Dorset: Robinson and Associates)
- Hoppe V 1985 Faith and fact in finite elements. *Finite Element News* 2: 12–15
- Huang H C, Hinton E 1984 A nine-code Lagrangian Mindlin plate element with enhanced shear interpolation. *Eng. Comput.* 1: 369–379
- Hughes T J R 1987 *The finite element method: Linear static and dynamic finite element analysis* (Englewood Cliffs, NJ: Prentice Hall)
- Hughes T J R, Cohen M, Haroun M 1978 Reduced and selective integration techniques in the finite element analysis. *Nucl. Eng. Design* 46: 206–222
- Hughes T J R, Taylor R L, Kanoknukulchal W 1977 A simple and efficient finite element for plate bending. *Int. J. Numer. Methods Eng.* 11: 1529–1543
- Jang J, Pinsky P M 1988 Convergence of curved shell elements based on assumed covariant strain interpolations. *Int. J. Numer. Methods Eng.* 26: 329–347
- MacNeal R H 1982 Derivation of element stiffness matrices by assumed strain distributions. *Nucl. Eng. Design* 70: 3–12
- Malkus D S, Hughes T J R 1978 Mixed finite element method – reduced and selective integration techniques: unification of concepts. *Comput. Methods Appl. Mech. Eng.* 15: 68–81
- Mohr G A 1982 Application of penalty functions to a curved isoparametric axisymmetric thick shell element. *Comput. Struct.* 15: 685–690
- Naganarayana B P 1991 *Consistency and correctness principles in quadratic displacement type finite element*. Ph D thesis, Indian Institute of Science, Bangalore
- Naganarayana B P, Prathap G 1989a Displacement and stress predictions from field- and line-consistent versions of the 8-node Mindlin plate element. *Comput. Struct.* 33: 1095–1106
- Naganarayana B P, Prathap G 1989b Force and moment corrections for the warped four-node quadrilateral plane shell element. *Comput. Struct.* 33: 1107–1115
- Naganarayana B P, Prathap G 1991 Field consistency analysis of 27-noded hexahedral elements for constrained media elasticity. *Finite Elements Anal. Design* 9: 149–168
- Naganarayana B P, Prathap G, Dattaguru B, Ramamurthy T S 1992 A field-consistent and variationally correct representation of transverse shear strains in the nine-noded plate element. *Comput. Methods Appl. Mech. Eng.* 97: 355–374
- Noor A K, Anderson C M 1982 Mixed models and reduced-selective integration displacement models for nonlinear shell analysis. *Int. J. Numer. Methods Eng.* 18: 1429–1454
- Noor A K, Hartley S J 1977 Nonlinear shell analysis via mixed isoparametric elements. *Comput. Struct.* 7: 615–626
- Noor A K, Peters J M 1981 Mixed models and reduced-selective integration displacement models for nonlinear analysis of curved beams. *Int. J. Numer. Methods Eng.* 17: 615–631
- Ojalvo I U 1974 Improved thermal stress by finite element methods. *AIAA J.* 12: 1131–1132

- Oleson J F 1983 Field-redistribution in finite elements – a mathematical alternative to reduced integration. *Comput. Struct.* 17: 157–159
- Park K C 1984 Symbolic Fourier analysis procedures for C-O finite elements. *Innovative methods for nonlinear analysis*, pp. 269–293
- Pawsey S E, Clough R W 1971 Improved numerical integration of thick shell finite elements. *Int. J. Numer. Methods Eng.* 3: 545–586
- Pittr J, Hartl H 1980 Improved stress evaluation under thermal load for simple finite elements. *Int. J. Numer. Methods Eng.* 15: 1507–1515
- Prathap G 1993 *Finite element method in structural mechanics* (Dordrecht: Kluwer)
- Prathap G 1985a A  $C^0$  continuous four-noded cylindrical shell element. *Comput. Struct.* 21: 995–999
- Prathap G 1985b The curved beam/deep arch/finite ring element re-visited. *Int. J. Numer. Methods Eng.* 21: 389–407
- Prathap G 1985c The poor bending response of the four-noded plane stress quadrilateral element. *Int. J. Numer. Methods Eng.* 21: 825–835
- Prathap G 1986 Field-consistency – Toward a science of constrained multi-strain-field finite element formulations. *Sadhana* 9: 345–353
- Prathap G 1988 The variational basis for least squares field-redistribution of strain functions in the finite element formulations of constrained media elasticity. TM ST 8801, National Aeronautical Laboratory, Bangalore
- Prathap G G 1994 Locking, rank and non-singularity of penalty linked stiffness matrix and consistency of strain field *Comput. Struct.* (to be published)
- Prathap G, Babu C R 1986a An isoparametric quadratic thick curved beam element. *Int. J. Numer. Methods Eng.* 23: 1583–1600
- Prathap G, Babu C R 1986b Field-consistent strain interpolations for the quadratic shear flexible beam element. *Int. J. Numer. Methods Eng.* 23: 1973–1984
- Prathap G, Babu C R 1987 Field-consistency and violent stress oscillations in the finite element method. *Int. J. Numer. Methods Eng.* 24: 2017–2033
- Prathap G, Bhashyam G R 1982 Reduced integration and the shear-flexible beam element. *Int. J. Numer. Methods Eng.* 18: 195–210
- Prathap G, Dattaguru B, Naganarayana B P 1989 Development of general purpose finite element package for structural analysis of isotropic, anisotropic and layered-composite structures: Project Proposal to AR & DB. PP ST 8907, National Aeronautical Laboratory, Bangalore
- Prathap G, Naganarayana B P 1988 Consistency and orthogonality requirements and accurate stress recovery for the assumed strain eight-noded quadrilateral plate bending elements. TM ST 8805, National Aeronautical Laboratory, Bangalore
- Prathap G, Naganarayana B P 1990a Consistent force resultant distributions in displacement elements with varying sectional properties. *Int. J. Numer. Methods Eng.* 29: 775–783
- Prathap G, Naganarayana B P 1992 Field-consistency rules for a 3-noded shear-flexible beam elements under non-uniform isoparametric mapping. *Int. J. Numer. Methods Eng.* 33: 649–664
- Prathap G, Naganarayana B P 1994 Consistent thermal stress evaluation in finite elements (to be published)
- Prathap G, Naganarayana B P, Somashekar B R 1988 Field-consistency analysis of the isoparametric eight-noded plate bending element. *Comput. Struct.* 29: 857–873
- Prathap G, Nirmala K 1990 Finite element formulations of constrained media elasticity – A bibliography. *Finite Elements Anal. Design* 7: 253–270
- Prathap G, Somashekar B R 1988 Field- and Edge- consistency synthesis of a four-noded quadrilateral plate bending element. *Int. J. Numer. Methods Eng.* 26: 1693–1708
- Prathap G, Subramanian G, Babu C R 1986 Stress oscillations in plane stress modelling of flexure – a field-consistency approach. *Int. J. Numer. Methods Eng.* 24: 711–724
- Prathap G, Viswanath S 1983 An optimally integrated four-noded quadrilateral plate bending element. *Int. J. Numer. Methods Eng.* 19: 831–840
- Simo J C, Hughes T J R 1986 On the variational foundations of assumed strain methods. *J. Appl. Mech.* 53: 51–54
- Stolarski H, Belytschko T 1981 Membrane locking and reduced integration for curved elements. *J. Appl. Mech.* 49: 172–178

- Stolarski H, Carpenter N, Belytschko T 1985 A Kirchhoff-mode method for C-O bilinear and serendipity plate elements. *Comput. Methods Appl. Mech. Eng.* 50: 121-145
- Tang L, Chen W, Liu Y 1984 Formulation of quasi-conforming element and Hu-Washizu principle. *Comput. Struct.* 19(1-2): 247-250
- Tessler A, Dong S B 1981 On a hierarchy of conforming Timoshenko beam elements. *Comput. Struct.* 14: 335-344
- Walz J E, Fulton R E, Cyrus N J, Eppink R T 1970 Accuracy of finite element approximations to structural problems. NASA TN-D 5728
- Wilson E L 1970 SAP - A general structural analysis program. SESM Report 70-20, Dept. of Civil Engineering, University of California, Berkeley
- Wilson E L, Taylor R L, Doherty W P, Ghabussi T 1973 Incompatible displacement models. *Numer. Comput. Methods in Struct. Mech.* (ed. S T Ferves *et al*), Academic Press, pp. 43-57
- Zienkiewicz O C 1977 *The finite element method* 3rd edn (London: McGraw Hill)
- Zienkiewicz O C, Hinton E 1976 Reduced integration, function smoothing and non-conformity in finite element analysis. *J. Franklin Inst.* 302(56): 443-461
- Zienkiewicz O C, Taylor R L, Too J M 1971 Reduced integration techniques in general analysis of plates and shells. *Int. J. Numer. Methods Eng.* 3: 275-290



## **Distortion, degeneracy and rezoning in finite elements – A survey**

RAVIPRAKASH R SALAGAME\* and ASHOK D BELEGUNDU

The Department of Mechanical Engineering, The Pennsylvania State University, University Park, PA 16802, USA

\*E-mail: r3s@ecl.psu.edu

**Abstract.** The results obtained from finite element analysis are significantly affected by the quality of elements. In certain applications like shape optimization, crash analysis, metal forming, fluid flow analysis, and large displacement analysis, the finite element mesh is systematically updated in an iterative process. In such situations, in spite of an ideal starting mesh, the quality of elements could deteriorate, causing severely distorted elements. In extreme cases, the elements become degenerate and further progress of analysis is restricted. An understanding of the methods of quantifying element distortion helps in identifying 'bad' geometry and in deciding when to remesh. Knowledge about geometric configurations which cause degeneracy assists in controlling degeneracy during the analysis. This paper contains a survey of available distortion measures and degeneracy conditions for various elements in two and three dimensions. It is a review of the literature in this field in the last two decades. A brief review of rezoning is also included, since it is one of the more popularly used methods to correct a distorted mesh.

**Keywords.** Distortion; degeneracy; finite element; Jacobian; mapping; rezoning; grid optimization.

### **1. Introduction**

Distortion measures are usually functions of the coordinates of the element. Examples of distortion measures for four-noded quadrilaterals are aspect ratio, skew and tapers. Most of the available literature in defining distortion is for two-dimensional elements. These measures either use terms of the Jacobian matrix or simply define measures in terms of linear dimensions and angles. These are purely a priori, in the sense that they just capture the shape of the element quantitatively. Some authors have also studied the relationship between stiffness matrix terms and element shapes. The parameters used are condition number and trace of the stiffness matrix. It should be noted that the distortion measures only tell the user how distorted the element shape is with respect to a standard element. Whether this is good, bad or not relevant



depends on the problem at hand. The effect of distortion on errors cannot be assessed a priori. However, bad elements are usually sources of trouble especially in regions where the gradient of the solution parameter is high. For example, in elasticity problems, they are the regions of stress concentration.

A degenerate element is one for which the coordinate transformation becomes mathematically invalid. Our discussion is limited to isoparametric elements which use serendipity functions. The element in physical space,  $x, y, z$  is mapped to a standard element in  $\xi, \eta, \zeta$  coordinate system. For a one to one mapping, the Jacobian of the transformation matrix should not change sign or become zero anywhere in the element domain. The term 'Jacobian' denotes the determinant of the Jacobian matrix. Thus,  $J = \det [J]$ . The effect of the Jacobian being zero is discussed later in the paper. Degeneracy has to be avoided for a reliable solution. Detecting degeneracy mathematically is a problem of finding the zero of a polynomial. Simple thumb rules are possible only for linear elements and simplified higher order elements. For practical problems in engineering, the Jacobian can be sampled along the boundary and at Gauss points to identify degeneracy quickly. Once the mesh is identified to be bad for further analysis, it is modified using rezoning or adaptive techniques. Rezoning using Laplacian smoothing is the fastest and simplest way of modifying the mesh. Hence it is the most widely used rezoning method with different forms of weights in the rezoning formula. An adaptive strategy based on rezoning uses error measures as weights and iteratively rezones the mesh till convergence. Grid optimization is a more rigorous method of obtaining the optimum mesh for a given problem. The idea of grid optimization is to seek a minimum potential energy configuration with nodal coordinates as design variables along with nodal displacements.

This paper contains mainly two sections. The first section is a survey of various distortion measures proposed in the literature. It also includes the methods of element evaluation used to study the behaviour of elements to known displacement fields. The second section is a review of mathematical literature on nonvanishing of Jacobian in the element. A brief review of rezoning follows this section.

## 2. Distortion measures and element evaluation

The first step towards defining distortion measures is to parametrize the shape of the finite element in terms of quantities which are independent of coordinate systems. The coordinates of the element itself can constitute a set of parameters, but in that case, the distortion measure will be dependent on the coordinate system. In addition, it will be very difficult to make any intuitive judgement about the shape of the element from its coordinates. Hence in order to capture shape variation, independent of coordinate systems, various parameters have been proposed for linear and quadratic elements. Of these, the most popular are the ones for the four-noded bilinear isoparametric element. There are four shape parameters for this element—aspect ratio, skew, and two tapers along each coordinate direction. Using these four parameters, it can be shown that all types of shape changes can be independently captured. General guidelines are available to limit these values for any analysis. For example, most commercial codes recommend an aspect ratio between 1 and 3 (e.g., I-DEAS). However, distortion of the element is caused by a combination of all these parameters. Hence limits on these quantities independent of each other is not sufficient. This fact is another argument for the need to have a distortion measure which is representative

of the behaviour of the element. Most of these measures are discussed with 2-D examples like bilinear and quadratic elements. Some of the measures like distortion metric can be extended to 3-D elements. The literature available for 3-D distortion measures is very limited.

## 2.1 Shape parameters

Robinson (1985) expressed shape parameters in terms of simple polynomial coefficients with a clear physical meaning. These parameters are evaluated from the elements of the Jacobian matrix. This concept is reviewed here. In the finite element analysis of general structures, quadrilateral elements can be used on curved surfaces. The plane in which the element lies need not necessarily be flat. The stiffness matrix for a warped element is usually based on a flat projected plane. It is thus necessary to consider warpage as one of the parameters. The warped quadrilateral (figure 1) with straight edges is denoted by its corner nodes A, B, C, D. The reference nodes for the projected plane are denoted as 1 to 8. This plane contains the midpoints of each side. The warpage of a quadrilateral is measured by its deviation from a flat projected plane. Each corner node of the warped element will be at a distance  $h$  from the corresponding corner point of the projected plane (figure 1). The height  $h$  is given by  $h = \text{abs}(\vec{OA} \cdot \vec{Z})$ . For a flat element  $h = 0$ . The warpage is measured by the warpage parameter  $\theta$  (degrees) which is defined as

$$\theta = \sin^{-1}(h/l)$$

where  $l$  is half of the smallest side of the element. Shape parameters for four-noded flat (projected) quadrilateral elements can be written from the interpolation functions found in any text book on finite element analysis (e.g. see Chandrupatla & Belegundu 1991),

$$x(\text{or } y) = \sum_{i=1}^4 N_i x_i(\text{or } y_i)$$

where  $N_i$  are the standard shape functions for a four-noded quadrilateral element. The above equations show that eight parameters are needed to define a quadrilateral. In this case they are  $x$  and  $y$  coordinates of the four corner nodes. This form, however hides the significance of shape parameters. An alternative form of shape

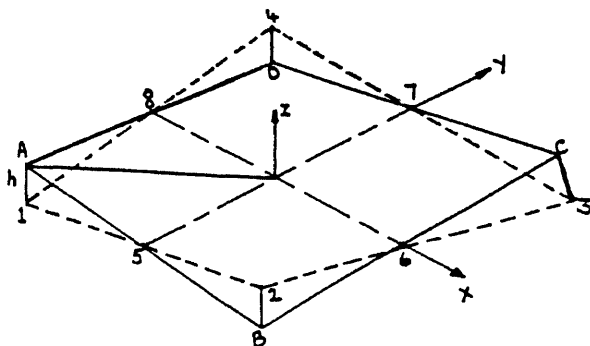


Figure 1. A warped quadrilateral element and its projected plane.

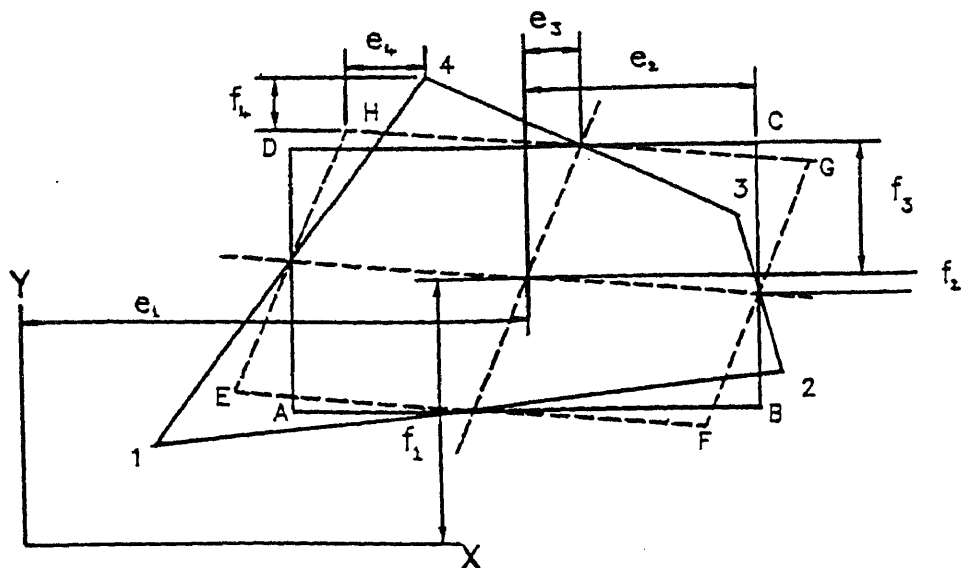


Figure 2. Shape parameters for a 4-noded quadrilateral.

representation is the polynomial form given by

$$\begin{aligned}x &= e_1 + e_2\xi + e_3\eta + e_4\xi\eta, \\y &= f_1 + f_2\xi + f_3\eta + f_4\xi\eta,\end{aligned}\quad (1)$$

where  $e$  and  $f$  are coefficients which can be related to the nodal coordinates by substituting the known values of  $x$  and  $y$  at the corner nodes. For example,  $e_1$  and  $f_1$  are given by

$$e_1 = (1/4)(x_1 + x_2 + x_3 + x_4) \quad \text{and} \quad f_1 = (1/4)(y_1 + y_2 + y_3 + y_4).$$

The physical significance of  $e_i$  and  $f_i$  are shown in figure 2, where 1-2-3-4 is the quadrilateral element. The rectangle  $ABCD$  is drawn through midpoints of the sides of the element so that the sides of the rectangle are parallel to the  $x$  and  $y$  coordinate axes. The dotted lines  $E-F-G-H$  form the parallelogram through midpoints of the element sides with its edges parallel to the  $\xi, \eta$  axes. The coefficients  $e_1$  and  $f_1$  define the coordinates of the geometric centre of the element,  $e_2$  and  $f_3$  define the size of the rectangle  $A-B-C-D$ ,  $e_3$  and  $f_2$  give two rotations (skew and rotation of the axes) and  $e_4$  and  $f_4$  give two tapers. If the local axes are defined with the origin at the centre of the element so that  $e_1 = f_1 = f_2 = 0$ , the required shape parameters can be defined as follows,

$$\text{Aspect ratio} = \text{Max}(e_2/f_3, f_3), \quad \text{Skew} = e_3/f_3,$$

$$\text{Taper along } x = f_4/f_3 \quad \text{and} \quad \text{taper along } y = e_4/e_2.$$

If the transformation from  $\xi, \eta$  to  $x, y$  axes is regarded as a continuous deformation, the aspect ratio represents element stretching and skew is equal to the shearing strain at  $\xi = 0, \eta = 0$ .

The Jacobian matrix for a flat (projected) quadrilateral can also be expressed in terms of the shape parameters. Specifically, the following expression for the determinant of the Jacobian matrix can be easily derived.

$$\det[J] = f_3^2 A (1 + T_x \xi + (T_y - (S/A) T_x) \eta),$$

where  $A$  is the aspect ratio,  $S$  is the skew and  $T_x$  and  $T_y$  are the tapers along  $x$  and  $y$  directions, respectively. The parameter  $f_3$  is a half side length of the basic rectangle (see figure 2). The coefficients in the Jacobian are therefore a function of the shape parameters.

**2.1a Shape parameters for an eight-noded quadrilateral:** For an eight-noded quadrilateral, a similar procedure to that for a four-noded quadrilateral can be used, but additional parameters come into the picture. These parameters account for the curved sides and midside nodes. Similar to (1), shape functions for an eight-node quadrilateral can be expressed in polynomial form (Robinson 1988)

$$x = e_1 + e_2 \xi + e_3 \eta + e_4 \xi \eta + e_5 \xi^2 + e_6 \eta^2 + e_7 \xi^2 \eta + e_8 \xi \eta^2,$$

$$y = f_1 + f_2 \xi + f_3 \eta + f_4 \xi \eta + f_5 \xi^2 + f_6 \eta^2 + f_7 \xi^2 \eta + f_8 \xi \eta^2,$$

where the coefficients  $e_i$  and  $f_i$  can be expressed in terms of the coordinates of the element. Referring to figure 3, the shape parameters are defined as given below. In figure 3 the local coordinate system is defined in such a way that  $e_1 = f_1 = f_2 = 0$ .

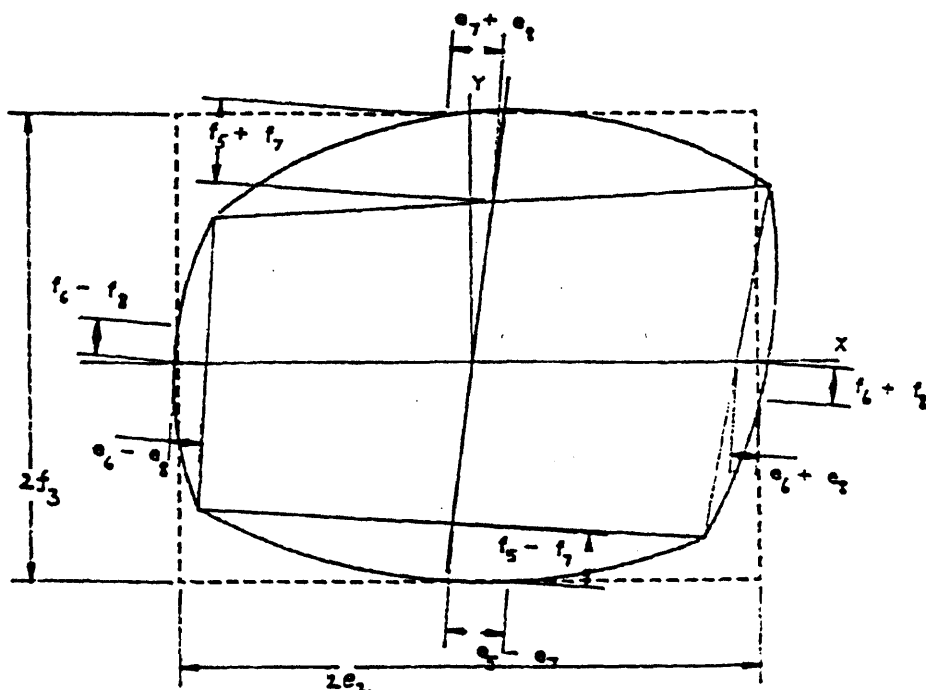


Figure 3. Shape parameters for an 8-noded quadrilateral.

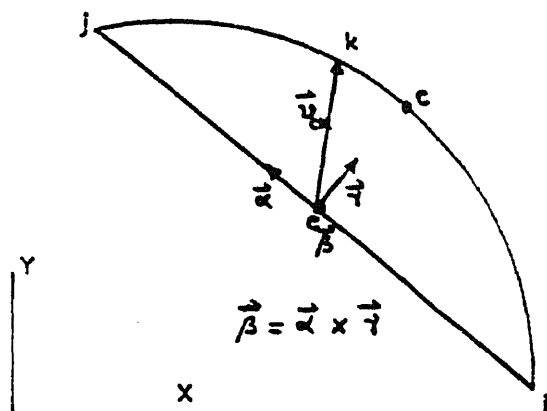


Figure 4. Tangential and normal deviations.

Aspect ratio =  $e_2/f_3$ , or  $f_3/e_2$  (larger of the two), Skew =  $e_3/f_3 + f_4/e_2$ ,  
taper along  $x = f_4/e_3$  and taper along  $y = e_4/e_2$ .

The other shape parameters describe the curvature of the sides and offset of midnodes. The offset of the node on a curved boundary from the midpoint of the associated chord is shown in figure 4. The offset measures are defined as the normal and tangential deviations.

Tangential deviation,  $TD = \alpha/(1/2) L$

where  $\alpha = \mathbf{v}_{ck} \cdot \boldsymbol{\alpha}$  is a unit vector along  $ij$ .

Normal deviation,  $ND = \gamma/(1/2) L$

where  $\gamma = \boldsymbol{\alpha} \times \boldsymbol{\beta}$ . The vector  $\boldsymbol{\beta}$  is a unit vector perpendicular to the plane containing the nodes  $i, j$  and  $k$ . The offset parameters are normalized with respect to half the chord length. When  $TD$  is zero, the node on the curved boundary is in the centre at  $C$  (see figure 4), and when  $ND$  is zero, the actual boundary is straight. A different way of defining distortion parameters for eight-noded quadrilaterals can be found in Hellen (1984). These are briefly described below.

Let  $\mathbf{r}$  denote the position vector of a point in the  $x, y$  coordinate system and  $\mathbf{r}_{,\xi}$  and  $\mathbf{r}_{,\eta}$  be the tangent vectors at any point  $\xi, \eta$ . Let  $\delta$  be the angle between the two tangent vectors. The distortion parameters are defined as

$$\text{Aspect ratio} = \frac{|\mathbf{r}_{,\xi}|}{|\mathbf{r}_{,\eta}|} \quad \text{or} \quad \frac{|\mathbf{r}_{,\eta}|}{|\mathbf{r}_{,\xi}|} \quad (\text{larger of the two}),$$

$$\text{skew} = 90^\circ - \delta.$$

Max. chord aspect ratio = (maximum chord length)/(minimum chord length),

Boundary node offset,  $R = (\text{length } ck)/(\text{length } ij)$ .

These parameters have been implemented in the BERQUAL program (Hellen 1984). The effect of these parameters on displacements for some structural problems and the discussion of the program can be found in Hellen (1986, 1987).

## 2.2 Shape sensitivity analysis

The basic idea of shape sensitivity studies is to vary the element shape parameters singly and in combination to assess their effect on the accuracy of the results produced, in a prescribed loading situation. This type of study have been carried out by Burrows (1986) and Robinson (1985, 1990) among others. In Burrows' work, a standard 2-D plane stress element is distorted in a predetermined way and subjected to standard loading conditions. Two types of loadings are considered – a constant stress and a constant moment. Theoretical nodal displacements and nodal forces are compared with FE nodal displacements and forces. Another method of element evaluation introduced by Robinson (1987) called continuum region element (CRE) testing is based on the idea that a rectangular continuum is available for which theoretical solutions are known for various loading conditions. Then the region is treated as a single element or a patch of elements. The applied load is in the form of specified displacements which are nonzero. The strain energy of the field due to FE formulation is computed and compared with the theoretical value of the strain energy. Examples on this approach to element testing can be found in the works of Robinson (1976, 1987, 1990). An analytical investigation on aspect ratio sensitivity of elements can be found in Robinson (1991). The elements studied are taken from commercial finite element codes like ABAQUS, MSC/NASTRAN, and ANSYS.

## 2.3 Distortion measure based on polynomial order of the element geometry

The shape parameters defined above for flat elements are easy to evaluate and implement in commercial codes. However there is need to understand their effects on solution errors. Since the error is a result of the combination of distortion and the variation of the unknown solution, a general correlation of the two is impossible. However, solution errors can be computed for assumed polynomial strain fields. This error can then be related to a known element distortion. Analytical study of distortion based on this concept was done by Barlow (1987, 1989). This approach for element evaluation is summarized below.

The basis of distortion measure is the polynomial order of the element geometry. Here, an example of this measure is presented with 8-noded isoparametric elements. The geometry of the element in an  $x, y$  coordinate system is written in a form which incorporates the distortion parameters. The form is,

$$[x \ y] = l[\xi \ \eta] + lF[\epsilon_x \ \epsilon_y],$$

where

$$F = [1, \xi, \eta, \xi^2, \xi\eta, \eta^2, \xi^2\eta, \xi\eta^2], \quad (2)$$

is a  $(1 \times 8)$  vector of polynomial terms in  $N$ ,  $4l^2 = \text{area of the element}$  (a scalar which makes the problem nondimensional.) and  $\epsilon_x$  and  $\epsilon_y$  are  $(8 \times 1)$  vectors of the distortion parameters given by

$$\epsilon_x^T = [\epsilon_{x1}, \epsilon_{x2}, \dots, \epsilon_{x8}],$$

$$\epsilon_y^T = [\epsilon_{y1}, \epsilon_{y2}, \dots, \epsilon_{y8}].$$

The first right-hand side term in (2) is the undistorted parent element and the remaining terms are the distortions of that element. These can be grouped into polynomial orders of distortion as given below.

The terms  $\varepsilon_{x1}$  and  $\varepsilon_{y1}$  represent just an offset from the origin. The terms  $\varepsilon_{x2}$ ,  $\varepsilon_{y2}$ ,  $\varepsilon_{x3}$  and  $\varepsilon_{y3}$  represent linear geometric distortions. They are similar to aspect ratio and skew parameters. The next three terms  $\varepsilon_{x4}$ ,  $\varepsilon_{y4}$ ,  $\varepsilon_{x5}$ ,  $\varepsilon_{y5}$ ,  $\varepsilon_{x6}$  and  $\varepsilon_{y6}$  are the quadratic distortions. These can be interpreted as anti-symmetric edge curvature, taper and symmetric midside node offset. The last two terms  $\varepsilon_{x7}$ ,  $\varepsilon_{y7}$ ,  $\varepsilon_{x8}$  and  $\varepsilon_{y8}$  are the cubic distortions. They represent symmetric edge curvature and anti-symmetric midside node offset. The values of the distortion parameters may be extracted from the nodal geometry by substituting nodal coordinates in (2) and solving for  $\varepsilon_x$  and  $\varepsilon_y$ . A local coordinate system must be defined at the geometric centre of the element to avoid 'pseudo' distortion measures due to the orientation of the element. This ensures that the distortions are the same for a given shape independent of the element orientation. For graphical description of this measure, see Barlow (1989).

**2.3a Element evaluation:** For assumed distortion, the element is evaluated based on two considerations – its ability to reproduce the required strain fields and the accuracy of integration of the strains to produce the element stiffness, i.e. numerical integration used to compute  $K$ . It is assumed that the nodal loads produce an equilibrium stress field of a given polynomial order. For a constant stress strain relationship, this would mean a similar strain field. Now if the elements are capable of reproducing the individual strain terms, it would also be capable of reproducing any linear combination, including those which represent the equilibrium strain field. Thus the following procedure is used in the evaluation.

A polynomial displacement field is assumed and the exact strains are obtained by differentiation. The same polynomial displacements are applied to the finite element and the strains are determined from the element displacement functions. The difference between two strain fields, the exact and the element representation, provides a measure of the error in the element strain representation. The coefficients of the polynomial terms in the applied displacement field will depend on element orientation. Hence the order of magnitude of the error is taken as the maximum which occurs in any individual polynomial term of a given degree. An example is given below. Consider the effect of distortion  $\varepsilon_{x3}$  (skew) on a cubic displacement field. Let  $u = x^3$ . Let the element have only skew parameter i.e.  $\varepsilon_{x3}$  and other parameters be zero. The  $x - y$  coordinate system is defined at the element centre with  $x$ -axis along the  $\xi$  axis. The applied strain field in terms of  $\xi, \eta$  is given by,

$$\varepsilon_{\text{exact}} = 3\xi^2 + 6\varepsilon_{x3}\xi\eta + 3\varepsilon_{x3}^2\eta^2,$$

while the finite element strain is given by,

$$\varepsilon_{FE} = 1 + 6\varepsilon_{x3}\xi\eta + 3\varepsilon_{x3}^2\eta^2.$$

The error in this case would be  $e = 3\xi^2 - 1$ . This evaluation procedure can be used for any element. In fact, Barlow (1989) has demonstrated this method for 20-noded brick elements in addition to quadrilateral elements. Based on this study, points at which minimum error occurs in the element domain can be obtained. In the above example,  $e$  is zero at Gauss points  $(\xi, \eta) = (\pm 1/\sqrt{3}, \pm 1/\sqrt{3})$ . These points are called optimal stress points. The procedure to analytically determine optimal points for assumed strain fields can be found in the works of Barlow (1976) and Budkowska (1991). For linear displacement fields the optimal points lie at  $\xi = 0, \eta = 0$ .

## 2.4 Distortion measures based on the Jacobian matrix

The multiple distortion parameters described above are obtained from the geometry of the element. Although simple to evaluate, their effect on the accuracy of the solution are interdependent. Thus, if a decision about element quality has to be made based on some numerical measure of the element geometry, a combination of all the shape parameters has to be used. In the previous sections, it was shown that the Jacobian can be expressed in terms of these shape parameters. Hence the value of the Jacobian in the element domain reveals very useful information about the element distortion. In this section some measures based on elements of the Jacobian matrix are described.

**2.4a Distortion parameter:** Distortion parameter (DP) was introduced by Nicolas & Ciritipitioglu (1977). For a general quadrilateral element it is defined as

$$DP = (4 \det[J]_{\min}/A) \text{ and for three dimensional elements it is defined as}$$

$$DP = (8 \det[J]_{\min}/V)$$

where  $A$  and  $V$  are respectively the area and the volume of the element. They are used to nondimensionalize the parameter. Above,  $\det[J]_{\min}$  refers to the minimum value of the Jacobian determinant in the element. For a parallelogram shaped four-noded quadrilateral element, DP is 1.0. For a four-noded element which degenerates to a triangle, DP is 0.0. The value of DP becomes negative at some point if the four-noded element becomes concave. Similar interpretations for 3-D and general higher order elements are not as obvious. One of the major disadvantages of this parameter is that it cannot detect extremely degenerate parallelograms. A rectangle, with extremely high aspect ratio would still have a  $DP = 1.0$  and so does a highly skewed parallelogram. Distortion of elements to such shapes are quite common in applications involving large changes in finite element mesh, like crash analysis, large displacement analysis or metal forming. A variation of DP, which is defined as the ratio of maximum value of Jacobian to its minimum value in the domain is sometimes used. This parameter also has all the above mentioned problems.

## 2.5 Decomposition of the Jacobian matrix

Separate measures of element orthogonality, aspect ratio, orientation and volume are computed by algebraic decomposition of the Jacobian matrix in this measure (Kerlick & Klopfer 1982). The decomposition is first done by splitting off that part of the matrix which gives relative orientation of the computational coordinate axes with respect to the physical axes. For example, for a 2-D element, the direction cosines of the  $\xi$  axis with respect to the  $x$ -axis is given by

$$\cos(\alpha) = (x_{,\xi}/x_{,\xi}^2 + y_{,\xi}^2)^{1/2}.$$

The components of  $J$  are expressed in terms of metric tensors,  $g_{ij} = x_i x_j + y_i y_j$ . The parameters orthogonality, volume, orientation and aspect ratios are now expressed in terms of the components  $g_{ij}$ . This method is described for finite difference analysis in Kerlick's work (Kerlick & Klopfer 1982). These separate measures cannot describe combination of stretching and shearing.



## 2.6 Distortion metric

The distortion metric proposed by Oddy *et al* (1988) is based on the observation that an analogy can be drawn between element distortion and strain. If a body with the shape of the element in the computational space  $\xi, \eta, \zeta$  were deformed until it had the shape of the element in physical space,  $x, y, z$ , then the elements with large distortions would experience the analogue of large strains. Hence for this "pseudo deformation", Green's strain is measured. If  $J_{ij}$  are the elements of Jacobian matrix, in  $n$ -dimensional space ( $n = 2$  or  $3$ ), the elements of  $J$  are first normalized to neutralize the effect of size. Thus

$$J'_{ij} = J_{ij}/|J|^{1/n}.$$

Let  $C_{ij} = J'_{ki}J'_{kj}$ . Using standard tensor notations, the deviatoric strains can be written as

$$e_{ij} = (1/2) \left( C_{ij} - \frac{1}{n} C_{kk} \delta_{ij} \right).$$

The second invariant of this tensor in 2-D or 3-D can be written as

$$J_2 = (1/8) [C_{ij}C_{ij} - (1/n)(C_{kk})^2]$$

Since the factor 8 serves to just scale the above quantity, the distortion metric simply defined as the expression within the brackets. This distortion metric is a function of the fourth power of the elements of a Jacobian matrix. It is also a function of position and not a constant over the entire domain. The authors use the largest value of the metric at Gauss points as the measure of distortion. However there is no basis for this selection. The comparison of various values of distortion metric with errors for selected examples with known theoretical solutions are also discussed in the above reference. A direct relation of a known value of distortion to some level of error a priori is not possible since the error in any analysis is highly dependent on the problem itself. The measures discussed here should therefore be used only as a guideline to detect excessive distortion of the element which is in general not desirable.

The transformation of the element from the parent to a local coordinate system can be thought of as a continuous deformation process in which the final deformed shape is the shape of the element in the physical  $x, y, z$  coordinate system. By expressing this "pseudo" deformations as functions of coordinates in the  $\xi, \eta, \zeta$  system, and assuming large strains, it can be shown that the above measure is directly related to the strain energy density of this pseudo deformation.

## 2.7 Other indicators of element quality

So far we have discussed distortion measures which are solely dependent on element geometry. There are some indicators, like condition number of stiffness matrix, studied by some authors as an overall indication of the quality of discretization. Distortion is just one of the factors which is reflected in these indicators along with interpolation function used, order of the element, etc.. Notable work in this area has been done (Melosh 1963; Khanna 1965; Fried 1971-73). In Fried (1972), the condition number of the stiffness matrix due to nonuniform meshes is studied. Bounds are derived on the

condition number and expressed in terms of the extremal eigenvalues of stiffness and mass matrices and discretization parameters of the mesh. The condition number is related to the size of the elements,  $h$ . Similar to condition number, trace of the element stiffness matrix has been used by Rigby & McNeice (1972) for evaluating their performance. It is suggested that elements with lower trace are likely to be better for general loading cases. Strain energy of the element has also been used instead of stiffness for comparison of two elements (Melosh 1963; Khanna 1965). Khanna's method compares two element stiffness matrices by calculating the difference in strain energies. A modification of this method and comparison of hexahedron elements based on strain energy can be found in the reference by Rigby & McNeice (1972).

The performance of the element for a given loading situation can be directly studied by using elements with known distortion values and comparing quantities like stress errors, strain errors, strain energies, energy norm of the error etc.. For example, in Salmon & Abel (1989), shape distortion effects on nine-noded quadrilateral membrane elements is studied. The shape parameters described above, aspect ratio, skew, and tapers are used to define distortion. Stress and strain errors are plotted for each type of shape distortion. A similar study on quadrilateral elements can be found in a more recent reference (Liu & Elmaraghy 1992). The exact energy error is used as a basis of comparison of meshes with distorted elements of varying degree of distortion. A generalized finite element evaluation procedure for evaluating two- and three-dimensional elements is described (Dow *et al* 1985). This procedure compares the strain energy content and the strain distribution of the finite element model to that of the continuum region it represents for well-defined strain states. Triangular and quadrilateral elements undergoing a series of progressive initial distortions are used for evaluation. From the results of this evaluation, the authors suggest an algorithm to predict maximum strain energy error as a function of initial geometry.

### 3. Element degeneracy conditions for isoparametric elements

In this section, the determinant of the Jacobian matrix is studied in detail. For different elements, the mathematical conditions which dictate zero of the Jacobian are explored. The term 'Jacobian' denotes the determinant of the Jacobian matrix. Thus,  $J = \det [J]$ .

Using the strain displacement relationship  $\epsilon = \mathbf{B}\mathbf{q}$ , where  $\mathbf{q}$  is a vector of nodal displacements, the strain can be expressed in the local  $\xi, \eta, \zeta$  coordinate system as

$$\epsilon = (1/J)\mathbf{G}(\xi, \eta, \zeta)\mathbf{q}$$

where the elements of the vector  $\mathbf{G}$  are functions of  $\xi, \eta, \zeta$ . The matrix  $\mathbf{B}$  relates strains to nodal displacements. It can be seen from the above equation that if  $J$  is zero at some point within the element, it implies that the strains are infinite or that we have an indeterminate formula. If such points occur at Gauss points used for numerical integration, the calculations cannot be continued. Even otherwise, we can expect the accuracy to deteriorate as such a point is approached (Jordan 1970). Mathematically this condition represents a nonunique mapping between the computational space  $(\xi, \eta)$  and the physical space  $(x, y)$  in a local region around the point where  $J$  is zero. This follows from the inverse function theorem (see appendix A). Since  $J$  is a continuous function, the above requirement implies that  $J$  should not change sign anywhere in the domain. We note the condition when  $J$  is zero at some point in the

element as degeneracy. In this section, the conditions which cause degeneracy are reviewed for various elements.

### 3.1 Triangular element

For a three-noded linear triangular (CST) element, the magnitude of  $J$  is equal to twice the area of the triangle. Thus degeneracy for this element occurs when the triangle collapses to a straight line.

Next consider a 6-noded higher order triangular element. In general, all three sides can be curved. A general relationship which describes degeneracy is too complicated to obtain and hence special cases are considered. A triangle with all three straight sides and one with two straight sides cover the majority of the applications. For such a triangle the "quarter-point" rule was first derived by Mitchell *et al* (1971).

- (1)  $J > 0$  for all points in the triangular element if  $(1/4) < k_1, k_2, k_3 < (3/4)$
- (2)  $J = 0$  at some point in the triangular element if  $k_i$  takes a value in either of the intervals  $(0, 1/4)$  or  $(3/4, 1)$  for any value of  $i$  where  $i = 1, 2, 3$ .

Similar rules are also derived by Jordan (1970) using vector analysis. In addition, if the triangle has at least two straight edges and one curved side, as shown in figure 5, it is shown that the node 5 on the curved side should be in the shaded region so that  $J > 0$ . This region is formed by the two lines which are parallel to the straight sides and pass through their quarter points.

If all the three sides are curved, a general rule to detect degeneracy is not available. But some specific rules can be obtained. Jordan (1970) provides a simple geometric check. Refer to figure 5. Draw the line from node 1 to node 6 and extend it by  $1/3$  of its length to point  $6'$ . Similarly with 2 to 5 to get  $5'$ . Rotate CCW around node 3 from  $6'$  to  $5'$ . If the rotation angle is  $180^\circ$  or more,  $J$  will vanish somewhere (the converse need not hold). This geometric interpretation is simply a statement of a scalar triple product formula for  $J$  obtained from vector analysis. For details see (Jordan 1970).

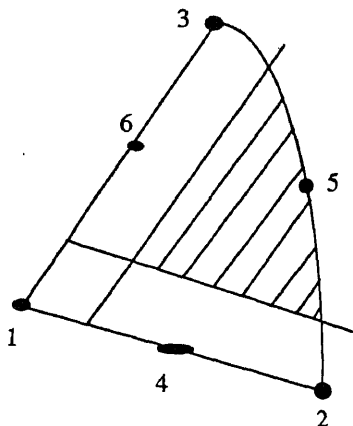


Figure 5. Valid region for midside node.

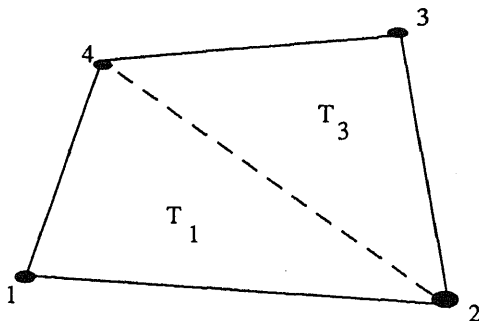


Figure 6. 4-noded quadrilateral element.

### 3.2 Four-noded quadrilateral element

The four-noded quadrilateral element is shown in figure 6. The shape functions are the serendipity functions most commonly used for this element (see Chandrupatla & Belegundu 1991). Using these functions, the Jacobian  $J$  can be expressed as

$$J = a + b\zeta + c\eta,$$

where  $a, b, c$  are functions of the nodal coordinates. Thus,  $J$  is a bilinear function of  $\xi, \eta$  and attains its minimum at the corners. The determinant of Jacobian has an interesting geometric interpretation for four-noded quadrilaterals (Okabe 1981). Refer to figure 6. Let  $T_i$  be the area of the partial triangle formed by the vertex  $i$  and two adjacent nodes. Then the Jacobian at node  $i$  can be written as

$$K_i = |J(\xi_i, \eta_i)| = (1/2) T_i, \quad i = 1, 2, 3, 4 \quad \text{and} \quad K_0 = |J(0, 0)| = (1/4) A,$$

where  $A$  is the area of the original quadrilateral. The coefficients  $a, b, c$  can be expressed in terms of  $K_i$  as

$$a = K_0,$$

$$b = \frac{1}{2}(K_2 - K_1),$$

$$c = \frac{1}{2}(2K_0 - K_1 - K_2).$$

Since  $J$  at the corners represent the area of the triangle described above,  $J = 0$  indicates that the quadrilateral degenerates to a triangle with the node at which  $J = 0$  being on the line joining its adjacent nodes. Hence a 4-noded quadrilateral element, which is not convex, has  $J = 0$  somewhere in its domain. The converse is also true. Thus, a four-noded quadrilateral element which has  $J = 0$  somewhere in the domain is nonconvex in shape.

### 3.3 Eight-noded quadrilateral

Detecting degeneracy and conditions that cause degeneracy in higher order elements is more complex. The most commonly used higher order quadrilateral element is an 8-noded quadrilateral (see figure 7).

The simplest form of this element is when its sides are straight lines. For this element, the restrictions on midside nodes were studied by Steinmueller (1974). The

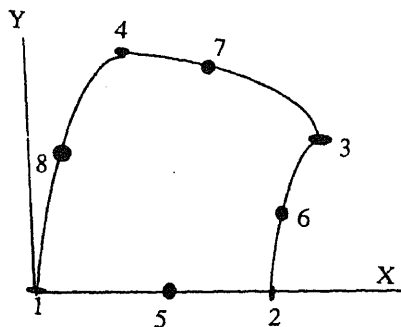


Figure 7. 8-noded quadrilateral element.

rules for positioning midside nodes is derived below. The approach given here is slightly simplified. In figure 7, assume that the origin of the coordinate system is at node 1 and the  $x$ -axis is along 1-2. Thus coordinates of 2 are  $(x_2, 0)$ . Let  $(a, 0)$  be the position of midside node 5. The  $x$ -coordinate of a point on 1-2 can be written as

$$x = N_2 x_2 + N_5 a = \frac{1}{2}(1 + \xi)\xi x_2 + (1 - \xi^2)a.$$

For invertible mapping i.e.  $J > 0$ , we need that  $J = (\partial x / \partial \xi) > 0$  from which we can show that

$$(1/4)x_2 < a < (3/4)x_2.$$

When the midside node is at any of these limiting positions, the Jacobian is zero at that point. Such elements are called 'quarter point elements'. They are widely used in fracture mechanics, since it was discovered that they possess the appropriate  $r^{1/2}$  singularity required in those problems. Numerical difficulties that may arise in 3-D quarter point elements due to negative Jacobian and aspect ratio are discussed by Peano and others (Peano 1987; Peano & Pasini 1982).

A geometrical check similar to the one described for a 6-noded triangle can also be used for 8-noded quadrilateral elements (Jordan 1970). However, for an element with all four curved sides, simple rules like quarter point rule, are not easily derived. A comprehensive mathematical analysis of a general 8-noded quadrilateral is presented (Field 1983). In this reference, algorithms have been developed to determine invertible transformations. The outline of the approach is given below.

Let  $U$  be the master element domain. The coordinates of any point  $(x, y)$  can be expressed using the shape function  $N_i$  as

$$x = \sum_{i=1}^8 N_i x_i \quad \text{and} \quad y = \sum_{i=1}^8 N_i y_i.$$

The Jacobian of the transformation matrix can be obtained from the above equation. The general form of the Jacobian can be written as

$$J(\xi, \eta) = k_0 \eta^3 + A(\xi) \eta^2 + B(\xi) \eta + C(\xi),$$

where  $k_0$  is a constant which depends on the element geometry,  $A$ ,  $B$  and  $C$  are the polynomials in  $\xi$ .  $A$  and  $B$  are quadratic and  $C$  is cubic. Thus, for a known value of  $\xi$ ,  $J$  is a cubic polynomial in  $\eta$ . Parallel expression can be written expressing  $J$  as a

cubic polynomial in  $\xi$  with variable coefficients in  $\eta$ ,

$$J(\xi, \eta) = v_0 \xi^3 + \alpha(\eta) \xi^2 + \beta(\eta) \xi + \gamma(\eta).$$

Since  $J$  is a continuous function on  $U$ , it must have a minimum and maximum value on  $U$ . The idea is to establish the sign of these values wherever they occur in the interior of  $U$ . The critical points of  $J(\xi, \eta)$  satisfy

$$\partial J / \partial \xi = \partial J / \partial \eta = 0, \quad \text{at } (\xi^*, \eta^*).$$

Thus

$$\partial J / \partial \xi = F(\xi, \eta) = A'(\xi) \eta^2 + B'(\xi) \eta + C'(\xi) = 0,$$

$$\partial J / \partial \eta = G(\xi, \eta) = 3k_0 \eta^2 + 2A(\xi) \eta + B(\xi) = 0. \quad (3)$$

The theory of resultant polynomials is used to identify critical points in  $U$ . A resultant of two polynomials is a polynomial whose coefficients depend upon the coefficients of the two given polynomials. The construction of resultants of two polynomials is described by Householder (1968). The ordinary resultant is a polynomial which has only the constant term. For the above two polynomials, the ordinary resultant is given by,

$$D(\xi) = \begin{bmatrix} A'(\xi) & B'(\xi) & C'(\xi) & 0 \\ 0 & A'(\xi) & B'(\xi) & C'(\xi) \\ 0 & 3k_0 & 2A(\xi) & B(\xi) \\ 3k_0 & 2A(\xi) & B(\xi) & 0 \end{bmatrix}.$$

From the theorem on resultants (Householder 1968), the necessary and sufficient condition for the two polynomials  $F$  and  $G$  above to have a common divisor of some degree greater than zero, is that the ordinary resultant  $D(\xi) = 0$ . A similar polynomial  $\delta(\eta)$  can be constructed by writing the polynomial equations in  $\xi$ . If both  $D$  and  $\delta$  are not identically zero, then the critical points of  $J(\xi, \eta)$  on  $U$  are isolated. In such a case, the equation  $D(\xi) = 0$  will have at most 7 real roots. For each of these roots, from the theorem of resultants, there will be atleast one  $\eta^*$  such that equation (3) is satisfied. The idea is to find these roots if they lie in  $U$ . Sturm sequence can be used to find the number of roots of  $D(\xi)$  that lie in  $[0, 1]$ .

The method described above is for a general case. Simplifications are possible for certain cases. The general algorithm for detecting invertible transformation can be summarized as below.

**Step 1:** Check if  $J > 0$  on the boundary. This is numerically straightforward since  $J(0, \eta)$ ,  $J(1, \eta)$ ,  $J(\xi, 0)$  and  $J(\xi, 1)$  are at most cubic polynomials in one variable. To show that any cubic polynomial  $P(r) > 0$ , show that  $P(0)P(1) > 0$  and that  $P'(r)$  has no roots in  $[0, 1]$ . If  $P(0)P(1) > 0$  and  $P'(c) = 0$  for  $c \in [0, 1]$ , then show that  $P(c) > 0$ .

**Step 2:** If  $J > 0$  on the boundary and if the element is a semiquadrilateral, i.e. one which has three straight edges and one curved side, then  $J$  is positive everywhere in  $[0, 1]$ . Stop. Otherwise go to step 3.

**Step 3:** If  $J > 0$  on the boundary of  $U$ , and if the function  $J$  reduces to  $J(\xi, \eta) = a_0 + a_1 \xi + a_2 \eta + a_3 \xi \eta$  then  $J$  is nonzero everywhere. Stop. Otherwise go to step 4.

**Step 4:** If  $J > 0$ , and  $k_0 = 0$ , evaluate  $F(\xi, -B(\xi)/2A(\xi))$ . If  $F \cong 0$ , then  $J$  is nonzero everywhere. Else use Sturm sequence to find if the roots lie in  $[0, 1]$  and if they do, check if  $J$  at those points is positive and stop. If  $k_0 \neq 0$ , go to step 5.

**Step 5:** If  $J > 0$  on the boundary, and if  $k_0 \neq 0$ , then use the method described above, i.e. solve  $D$  and then (6) to find the roots.

The above algorithm gives necessary conditions for evaluating the sign of  $J$  on  $U$ . The following are the sufficiency conditions for the nonvanishing of  $J$ . For proof of these conditions see (Field 1983).

If  $J > 0$  on the boundary of  $U$ ,

(A) if  $3k_0 + A(\xi) \leq 0, 0 < 0, 0 < \xi < 1$ , then  $J(\xi, \eta) > 0$  on  $U$ ,

(B)  $3k_0 + A(\xi) > 0$  and  $3k_0 + 2A(\xi) + B(\xi) < 0, 0 < \xi < 1$ , then  $J(\xi, \eta) > 0$  on  $U$ ,

(C) if  $k_0 \leq 0$ , then each of the following conditions imply  $J > 0$  on  $U$ .

$$B(\xi) + C(\xi) > 0, \quad 0 < \xi < 1,$$

$$\phi = \{3k_0 + 2A(\xi) + B(\xi) > 0\} \cap \{B(\xi) + C(\xi) < 0\}.$$

(D) if  $k_0 > 0$ , then each of the following imply  $J > 0$  on  $U$ ,

$$A(\xi) > 0, B(\xi) > 0, \quad 0 < \xi < 1,$$

$$A(\xi) \geq 0, B(\xi) \leq 0, B(\xi) + C(\xi) \geq 0, \quad 0 < \xi < 1.$$

Global inversion of bilinear and quadratic isoparametric mapping is also studied in detail by Frey *et al* (1978).

### 3.4 Other higher order quadrilateral elements

A general algorithm similar to the one derived for quadrilateral element is too unwieldy for higher order quadrilateral elements. Even for simple cases of these elements, the solution becomes very complex. Sometimes, graphical methods are used to find the conditions of degeneracy. For example, consider a 12-noded quadrilateral element (cubic) with all straight sides as shown in figure 8. For invertible mapping,

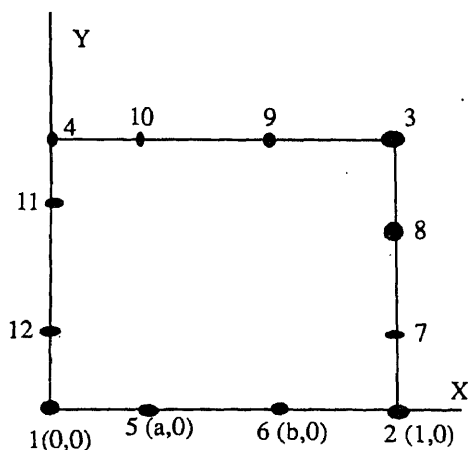


Figure 8. 12-noded quadrilateral element.

it can be shown that the following condition has to be satisfied.

$$(9/2)(2 - 10\xi - 9\xi^2)a + (9/2)(-1 + 8\xi - 9\xi^2)b + (1/2)(2 - 18\xi + 27\xi^2) > 0.$$

This inequality represents a region bounded by the envelope of an infinite number of straight lines, represented by the parameter,  $\xi$ ,  $0 \leq \xi \leq 1$ . By using various values of  $\xi$  and drawing these straight lines, we get a closed curve. Admissible values of  $a$  and  $b$  can be obtained from the interior of this curve.

Similar analysis for cubic isoparametric transformation of a nine-noded triangle is discussed in Mitchell (1979) and Woodford *et al* (1978). A triangle with two straight sides and one cubic side is chosen for this analysis. The effect of certain interpolating polynomials of degree four and five on triangular elements with one curved side is studied in Stephenson & Manohar (1979).

### 3.5 Isoparametric solids – tetrahedral elements

For a 4-noded tetrahedral element, which is the simplest solid element, determinant of Jacobian,  $J$ , is a constant and proportional to the volume of the element. The relationship is given by,  $|J| = 6V_e$ , where  $V_e$  is the elemental volume. Hence a zero-Jacobian would imply degeneration of tetrahedra to a triangle. For a higher order tetrahedral element, the Jacobian check algorithm described for 8-noded quadrilaterals can be extended easily (Field 1981). Consider a 10-noded tetrahedron element as an example. The Jacobian is of the form

$$J(\xi, \eta, \zeta) = k_1 \xi^3 + A(\eta, \zeta) \xi^2 + B(\eta, \zeta) \xi + C(\eta, \zeta),$$

where  $k_1$  is a constant and  $A$ ,  $B$  and  $C$  are given by

$$A(\eta, \zeta) = A_{10}\eta + A_{01}\zeta + A_{00},$$

$$B(\eta, \zeta) = B_{20}\eta^2 + B_{11}\eta\zeta + B_{02}\zeta^2 + B_{10}\eta + B_{01}\zeta + B_{00},$$

and

$$C(\eta, \zeta) = C_{30}\eta^3 + C_{21}\eta^2\zeta + C_{12}\eta\zeta^2 + C_{03}\zeta^3 + C_{20}\eta^2 + C_{11}\eta + C_{02}\zeta^2 + C_{10}\eta + C_{01}\zeta + C_{00}.$$

The constants in the above equation are found from nodal coordinates. The first step is to verify that  $J$  is positive on the boundary of the element. Since  $J$  is a function of two variables on each face of the element, the algorithm described for two variable functions under quadratic elements can be used. The next step is to check the sign of  $J$  on the interior of the element. Similar to results of 2-D, simplified results can be obtained under certain conditions. The following is an example. Let

$$L_1(\xi, \eta, \zeta) = 3k_1\xi + A_{10}\eta + A_{01}\zeta + A_{00},$$

$$L_2(\xi, \eta, \zeta) = B_{20}\xi + 3C_{30}\eta + C_{21}\zeta + C_{20},$$

$$L_3(\xi, \eta, \zeta) = B_{02}\xi + C_{12}\eta + C_{03}\zeta + C_{02}.$$

If  $J > 0$  on the boundary of the element and if one of the following conditions is satisfied, then  $J > 0$  everywhere.

$$(1) L_1(0, 0, 0) \leq 0, \quad L_1(0, 0, 1) \leq 0, \quad L_1(0, 1, 0) \leq 0 \quad \text{and} \quad L_1(1, 0, 0) \leq 0$$



or

$$(2) L_2(0,0,0) \leq 0, \quad L_2(0,0,1) \leq 0, \quad L_2(0,1,0) \leq 0 \quad \text{and} \quad L_2(1,0,0) \leq 0$$

or

$$(3) L_3(0,0,0) \leq 0, \quad L_3(0,0,1) \leq 0, \quad L_3(0,1,0) \leq 0 \quad \text{and} \quad L_3(1,0,0) \leq 0$$

For a 10-noded tetrahedron with three straight edges, each straight edge having a bisecting midside node and all three straight edges intersecting at  $(0,0,0)$ , the Jacobian check algorithm has been described by Field (1981).

### 3.6 Isoparametric solid elements – 8-noded brick element

For the 4-noded quadrilateral element, a positive  $J$  at the corners is an indication of convexity and positive  $J$  in the interior of the element. However a 8-noded brick need not be convex to have a positive Jacobian. Neither is a positive Jacobian at its corner nodes, a sufficient condition to assure that the associated transformation has positive Jacobian. For example, consider an element with coordinates  $(0, 1/4, 1)$ ,  $(1, 0, 1)$ ,  $(1, 1/4, 1)$ ,  $(0, 0, 0)$ ,  $(0, 3/4, 1)$ ,  $(1, 1, 1)$ ,  $(1, 3/4, 0)$  and  $(0, 1, 0)$ . The element is not convex, but its Jacobian is positive. Similarly an element with coordinates  $(0, 0, 1)$ ,  $(4, 0, -1)$ ,  $(1, 0, 0)$ ,  $(0, 0, 0)$ ,  $(0, 1, 1)$ ,  $(1, 1, 0)$ ,  $(1, -1, -1)$  and  $(0, 1, 0)$  has positive  $J$  at its corners, but  $J(1/2, 0, 0) = 0$ . The conditions to ensure positive  $J$  for the element whenever  $J$  is positive on its boundary has been derived (Field 1981). The procedure is exactly similar to the one above for tetrahedral elements. First, the non-negativity of the Jacobian has to be ensured on the boundary of the element. On each face the Jacobian is a polynomial in two variables and on each edge it is a quadratic in one variable. After verifying that it does not vanish somewhere on the edges of a face, verifying that it does not vanish on the interior of a face requires solution of

$$\frac{\partial J}{\partial \xi} = \frac{\partial J}{\partial \eta} = \frac{\partial J}{\partial \zeta} = 0.$$

Each of these derivatives is linear in the corresponding variable and on each face, one of the three variables is zero. The condition which implies a positive  $J$  throughout the element for a positive  $J$  on the entire boundary can be obtained using a similar approach as the tetrahedron. For details, see Field (1981). Degeneracy of 8-noded brick elements is also discussed by Peano (1987).

Alternatives to isoparametric transformation have also been considered. For example, Mitchell *et al* (1971) proposed a technique whereby interpolating functions are obtained directly in terms of  $x$  and  $y$  for the triangle and quadrilateral with arbitrary midside points. Wachspress (1971) developed a basis of wedge functions for convex polygon boundaries. Other notable efforts in this direction can be found in the works by Gordon & Hall (1973), McLeod (1976, 1978) and McLeod & Mitchell (1975). In spite of other alternatives, isoparametric transformations are the most widely used in displacement formulation of finite element analysis.

## 4. Rezoning and grid optimization methods

Rezoning, by definition, is simply relocation of the nodes based on some criteria. The criteria used could be as simple as improving the geometry of the element so that

they are less distorted or it could be based on the feedback from the finite element solution. In general, the process of rezoning involves two questions – when to rezone and how to rezone. The first question is generally answered by using well known measures of element quality which are purely geometrical in nature. These measures have been discussed in the previous chapter. The error measures based on stress jumps, or energy norms can also be used to decide when to rezone. The algorithm to rezone the mesh is usually based on some kind of nodal averaging which is a variation of the “Laplacian smoothing”.

The problem of obtaining an adequate finite element mesh for a given problem can also be formulated and solved as an optimization problem. An optimized mesh has been shown to be twice as efficient as an evenly divided mesh in terms of the number of degrees of freedom to produce the same accuracy (McNiece & Marcal 1973; Turcke & McNiece 1974). Details on grid optimization and other methods are reviewed in this section. This review of rezoning methods is by no means complete. However key references have been discussed as a guideline.

#### 4.1 Laplacian smoothing

This is the most popular method of grid smoothing, originally conceived by Winslow (1967). If a node  $n$  is shared by elements  $e_1, e_2, e_3, \dots, e_n$ , then its coordinates  $(x_n, y_n)$  are modified by

$$x_n = \frac{\sum_{i=1}^N x_{ci} w_i}{\sum_{i=1}^N w_i} \quad \text{and} \quad y_n = \frac{\sum_{i=1}^N y_{ci} w_i}{\sum_{i=1}^N w_i},$$

where  $x_{ci}$  and  $y_{ci}$  are the coordinates of the centroid of the element  $e_i$  and  $w_i$  can be considered as weights appropriately defined for averaging. If refined elements are needed in a certain region of the mesh, large weights have to be given in those regions in the corresponding elements. The solution to the above equation is a rough approximation of the discrete Laplacian. Thus the above formula is strictly valid for orthogonal meshes. Since the solution of Laplace's equation is known to have an averaging property, the above formula is used for rezoning any arbitrary mesh which is distorted. Instead of the nodal coordinates, as described above, the coordinates of the geometric centre of the element which is connected to the node are used. The weights are chosen as some parameter which is to be averaged over the domain. In the simplest form, we can use  $w_i = 1.0$ . The distortion measure is another choice, but its usefulness in reducing solution error has to be explored. Note that the above formula is used iteratively. Successive use of the formula over a number of iterations converges it to a more uniform mesh.

#### 4.2 Grid optimization methods

Grid optimization is a systematic way of obtaining finite element meshes that yield the required accuracy for the minimum effort. During early investigation of this problem (Felippa 1977), nodal coordinates were used as design variables in the appropriate energy functional. The resulting equations are highly nonlinear and need lot of computational effort. Hence attempts to find true optima are not very practical. Methods to obtain near-optimum grids based on the application of a solution-based criterion were proposed by Shephard *et al* (1980) and Turcke & McNiece (1974).

There are two key issues to be addressed in the application of these criteria methods – the criteria to be used, and the method of grid enrichment to be employed. The method of grid enrichment used might increase the degrees of freedom in the structure or just redistribute the nodes.

The requirements for optimum grids were studied by Turcke & McNeice (1974). These are based on the optimum grids obtained for various two dimensional problems by direct optimization using nodal coordinates as design variables along with nodal displacements. The optimization problem is described below.

The total potential energy obtained from finite element formulation can be written as

$$\Pi = \frac{1}{2} \mathbf{a}^T \mathbf{K} \mathbf{a} - \mathbf{a}^T \mathbf{F},$$

where  $\mathbf{a}$  is a column vector of unknown nodal displacements,  $\mathbf{K}$  is the stiffness matrix and  $\mathbf{F}$  is the column vector of nodal point forces. If we consider nodal coordinates  $x_i$  as design variables along with displacements, then minimum potential energy implies that

$$(\mathbf{K} \mathbf{a} - \mathbf{F}) = 0,$$

$$\frac{1}{2} \mathbf{a}^T \frac{\partial \mathbf{K}}{\partial x_i} \mathbf{a} - \frac{\partial \mathbf{F}}{\partial x_i} \mathbf{a}^T = 0.$$

Thus the feasible set of nodal co-ordinates and nodal displacements in the above system of nonlinear equations is one that ensures that the potential energy functional is stationary. A sufficient condition for the solution of the above equation to be a minimum is that the Hessian of  $\Pi$  be positive definite. Using the above relations, the best possible approximate solution for the given finite element mesh is obtained. The above optimization problem is solved using Rosenbrock's direct search technique. The authors (Turcke & McNiece 1974) provide specific guidelines for obtaining optimum grids. They suggest that the element edges have to be aligned along isoenergetics i.e. along contours of constant strain energy density. These contours can be obtained from the initial analysis on a coarse grid.

Furthermore, Shephard *et al* (1980) developed an interactive approach to synthesis of near-optimal meshes. The criteria used is variation of strain energy density in the element based on an argument that the best finite element solution is the one that best approximates the total strain energy. Hence this solution is associated with the best possible approximation to the integrand of the strain energy, i.e. SED. Key nodes are placed on the boundary of the mesh using above criteria and the new mesh is generated.

A variation of strain energy density has been used in Melosh & Marcal (1977) for mesh enrichment. The authors study the effect of gradual introduction of an additional degree of freedom on the potential energy. This effect is measured by 'specific energy difference', which is estimated by taking the difference between the SED at any point and that at the centroid of the element. A criteria based on a measure of the interpolation error associated with the finite element model is used by Kikuchi (1986). In this reference, the nodes are redistributed by an iterative scheme. More details on this method can be found from Diaz *et al* (1983). Mesh modification based on minimization of an interpolation error estimate is presented by Demcowicz & Oden (1986). The examples in this reference are taken from fluid flow problems. A more recent reference on mesh optimization problem (Martinez & Samartin 1991) presents closed form

solutions to determine optimal mesh for 1-D problems. Some practical examples of the use of rezoning and adaptive remeshing can be found in the works of Jung-Ho Cheng & Kikuchi (1986), Jung-Ho Cheng (1988), Liefvooghe & Fleury (1990), Yon & Yang (1990) and Yang *et al* (1989).

#### 4.3 General method for adaptive grid design

The optimal grid-design problem is defined by Kikuchi (1986)

Minimize (Maximum  $E_e$ ),

where  $E_e$  is an error measure of  $e$ th finite element. The necessary condition for the above optimization problem is that (Babuska *et al* 1983),

$E_e = \text{constant}, \quad e = 1 \text{ to no. of elements.}$

For stationary heat conduction problems and elastostatic problems, the smoothing scheme based on the Laplacian smoothing works well by taking the weights as  $E_e/A_e$  where  $A_e$  are element areas (Kikuchi 1986). For one-dimensional problems, the necessary condition is satisfied by repeated application. For two-dimensional problems, repeated application may not yield the necessary optimum, but it always reduces the maximum value of error at the beginning and converges to a fixed grid which is very close to the optimal one. If the initial mesh has only rectangular elements and if the necessary condition is satisfied, the application of Laplacian smoothing does not alter the mesh topology. If the mesh has elements of irregular shape, the smoothing scheme moves the grids even if the necessary condition is satisfied.

#### 4.4 Other approaches

Some authors have approached the problem purely from geometric considerations, i.e. optimization to minimize mesh distortion. For example, in a recent reference on mesh smoothing (Parthasarathy & Kodiyalam 1990), a constrained optimization problem is solved starting from a valid initial mesh. The objective function is chosen to be the RMS value of the aspect ratio of the elements. Kennon & Dulckravich (1986) use a linear combination of orthogonality and mesh smoothness measures as cost functions for optimization. Smoothness measure at any node is defined as the sum of the squares of the differences in area between adjacent elements connected to that node. The orthogonality measure at node  $i$  is defined as the sum of the squares of the dot products of the adjacent vectors which emanate from the node  $i$ . Minimizing the weighted sum of these measures ensures a smooth and orthogonal grid. This approach has been used in 3-D problems using volumes as measures in Carcaillet *et al* (1986). Similar work can also be found in Hayes *et al* (1986).

The problem of rezoning and grid optimization has been a subject of study for more than two decades now. A comprehensive review of the subject can be found in Shephard (1979). More recent papers have concentrated on applications of these methods to practical problems in engineering like flow analysis, metal forming, shape optimization etc. For iterative solution methods like large displacement analysis, the variables like displacement and stresses have to be transferred to the new positions of nodal points. A method to do this by inversion of isoparametric mapping is described in Crawford *et al* (1989).

## 5. Conclusions

It should be noted that solution errors are a result of the combination of factors – element distortion, the nature of the problem, degree of polynomial used for interpolation, and boundary conditions. Hence, a priori knowledge of the way distortion affects accuracy of analysis is impossible to predict. However, large distortions in general cause large errors. The distortion measures described in this review can be used to detect such distortions. Empirical rules for acceptable levels of distortion are quite common, though these are merely heuristic. Quantifying distortion using a reasonably good measure is very useful especially for rezoning or redefining the nodes to improve mesh quality.

Through the analysis of isoparametric elements for a nonvanishing Jacobian, it was observed that certain shapes of elements are 'forbidden' to maintain positive Jacobian in the element. The methods described here can be used to identify degeneracy.

For rezoning, weighted averaging at the nodes is the most popular method because of simplicity and ease of implementation. This scheme has certain disadvantages. The smoothness and uniformity of the rezoned mesh depends on the weighing factor used. With simple averaging, invalid geometry might result near boundaries for some problems. Though optimization seems to be a better approach, the computational effort can be very large for complex problems. Hence developing better rezoning techniques which yield less distorted and more accurate elements with computational ease is a subject of research.

## Appendix A

### *Theorem on inverse transformation*

Let  $x = f(u, v)$ ,  $y = g(u, v)$  define a continuously differentiable transformation for all pairs  $(u, v)$  in some neighbourhood of a point  $(u_0, v_0)$ . Let  $x_0 = f(u_0, v_0)$ ,  $y_0 = g(u_0, v_0)$ , and suppose that the Jacobian is not zero at  $(u_0, v_0)$ . Then there exist positive numbers  $a, b, \alpha, \beta$  and functions  $F(x, y)$ ,  $G(x, y)$  defined when  $|x - x_0| < a$ ,  $|y - y_0| < b$  such that the following assertions are true:

Let  $R$  be the rectangular region in the  $xy$ -plane defined by the inequalities  $|x - x_0| < a$ ,  $|y - y_0| < b$  and let  $S$  be the rectangular region in the  $uv$ -plane defined by the inequalities  $|u - u_0| < \alpha$ ,  $|v - v_0| < \beta$ . Then

(1) To any  $(x, y)$  in  $R$  corresponds a unique  $(u, v)$  in  $S$  such that  $x = f(u, v)$ ,  $y = g(u, v)$ , and this unique pair is given by

$$u = F(x, y), \quad v = G(x, y).$$

(2) The functions  $F$  and  $G$  are continuous and have continuous partial derivatives given by

$$\frac{\partial F}{\partial x} = \frac{1}{J} \frac{\partial g}{\partial v}, \quad \frac{\partial F}{\partial y} = -\frac{1}{J} \frac{\partial f}{\partial v}, \quad \frac{\partial G}{\partial x} = -\frac{1}{J} \frac{\partial g}{\partial u} \quad \text{and} \quad \frac{\partial G}{\partial y} = \frac{1}{J} \frac{\partial f}{\partial u},$$

where  $J = \frac{\partial(f, g)}{\partial(u, v)}$  and  $u, v$  are expressed in terms of  $x, y$ .

## References

- Babuska I, Chandra J, Flaherty J E 1983 *Adaptive computational methods for partial differential equations* (Philadelphia, PA: SIAM)
- Ball 1980 The interpolation function of a general serendipity rectangular element. *Int. J. Numer. Meth. Eng.* 15: 773-778
- Backlund J 1978 On isoparametric elements. *Int. J. Num. Meth. Eng.* 12: 731-732
- Barlow J 1976 Optimal stress locations in finite element models. *Int. J. Numer. Meth. Eng.* 10: 243-251
- Barlow J 1977 Comment on optimal stress locations in finite element models. *Int. J. Numer. Meth. Eng.* 11: 604
- Barlow J 1987 Distortion effects in isoparametric elements, an analytical approach. *National Agency for Finite Element Methods and Standards, AGM Technical Session*
- Barlow J 1989 More on optimal stress points-reduced integration, element distortions and error estimation. *Int. J. Numer. Meth. Eng.* 28: 1487-1504
- Budkowska B B, Fu Q 1991 Analytical determination of the optimal strain and stress points for the displacement model of the finite element method. *Comput. Struct.* 41: 937-946
- Burrows D J 1986 A finite element shape sensitivity study. *Reliability methods for engineering analysis* (eds) K J Bathe, D R J Owen (Pineridge Press)
- Carcaillet R, Dulikravich G S, Kennon S R 1986a Generation of solution-adaptive grids using optimization. *Comput. Meth. Appl. Mech. Eng.* 57: 279-295
- Carcaillet R, Kennon S R, Dulikravich G S 1986b Optimization of three-dimensional computational grids. *J. Aircraft* 23: 415-421
- Chandrupatla T R, Belegundu A D 1991 *Introduction to finite elements in engineering* (Englewood Cliffs, NJ: Prentice Hall)
- Clough R W 1969 Comparison of three-dimensional finite elements. *Proceedings of the symposium on application of finite element methods in civil engineering*, Vanderbilt University, Nashville, TN, pp 1-26
- Crawford R H, Anderson D C, Waggenspack W N 1989 Mesh rezoning of 2D isoparametric elements by inversion. *Int. J. Numer. Meth. Eng.* 28: 523-531
- Demkowicz L, Oden J T 1986 On a mesh optimization method based on a minimization of interpolation error. *Int. J. Eng. Sci.* 24: 55-68
- Diaz A R, Kikuchi N, Taylor J E 1983 A method of grid optimization for finite element methods. *Comput. Meth. Appl. Mech. Eng.* 41: 29-45
- Dow J O, Ho H T, Cabiness H D 1985 Generalized finite element evaluation procedure. *J. Struc. Eng.* 111: 435-452
- Felippa C A 1977 Optimization of finite element grids by direct energy search. *Appl. Math. Modelling* 1: 239-244
- Field D A 1981 An algorithm for determining invertible quadratic isoparametric finite element transformations. *Math. Comput.* 37: 347-360
- Field D A 1983 Determining invertible two and three dimensional quadratic isoparametric transformations. *Math. Comput. Simulation* 25: 485-488
- Frey A E, Hall C A, Porsching T A 1978 Some results on the global inversion of bilinear and quadratic isoparametric finite element transformations. *Math. Comput.* 32: 725-749
- Fried I 1971 Discretization and computational errors in high order finite elements. *AIAA J.* 9: 2071-2073
- Fried I 1972 Condition of finite element matrices generated from nonuniform meshes. *AIAA J.* 10: 219-221
- Fried I 1973 Accuracy and condition of curved (isoparametric) finite elements. *J. Sound Vibration* 31: 345-355
- Gifford L 1979 More on distorted isoparametric elements. *Int. J. Numer. Meth. Eng.* 14: 290-291
- Gordon W J, Hall C A 1973 Construction of curvilinear coordinate systems and applications of mesh generation. *Int. J. Numer. Meth. Eng.* 7: 461-477
- Hayes L J, Kennon S R, Dulikravich G S 1986 Grid orthogonalization for curvilinear alternating-direction techniques. *Comput. Meth. Appl. Mech. Eng.* 59: 141-154
- Hellen T K 1984 A finite element mesh quality assessment program. *Accuracy, Reliability and Training in FEM Technology*, Proceedings of the Fourth World Congress on Finite Element Methods, Interlaken, Switzerland

- Hellen T K 1987 Finite element quality assessment procedures. *Proceedings of NAFEMS International Conference Quality Assurance in Finite Element Analysis*, Brighton
- Hellen T K 1986 An approach to the quality assessment of finite element meshes. *Structural analysis systems* (ed.) A Niku-Lari (Oxford: Pergamon) vol. 5
- Householder A S 1968 Bigradients and the problem of routh and hurwitz. *SIAM Rev.* 10: 56-66
- Irons B 1968 Testing and assessing finite elements by an eigenvalue technique. *Proceedings of the conference on recent developments in stress analysis, new concepts and techniques and their practical application* (London: R. Aeronaut. Soc.) p. 12
- Jordan W B 1970 The plane isoparametric structural element. *AEC Research and Development Report KAPL-M-7112*
- Jung-Ho Cheng 1988 Automatic adaptive remeshing for finite element simulation of forming processes. *Int. J. Numer. Meth. Eng.* 26: 1-18
- Jung-Ho Cheng, Kikuchi N 1986 A mesh rezoning technique for finite element simulations of metal forming processes. *Int. J. Numer. Meth. Eng.* 23: 219-228
- Kennon S R, Dulikravich G S 1986 Generation of computational grids using optimization. *AIAA J.* 24: 1069-1073
- Kerlick G D, Klopfer G H 1982 Assessing the quality of curvilinear coordinate meshes by decomposing the Jacobian matrix. *Numerical grid generation* (ed.) J F Thompson, pp. 787-807
- Khanna J 1965 Criterion for selecting stiffness matrices. *AIAA J.* 3: 1976
- Khanna J, Hooley R F 1966 Comparison and evaluation of stiffness matrices. *AIAA J.* 4: 2105-2110
- Kikuchi N 1986 Adaptive grid-design methods for finite element analysis. *Comput. Meth. Appl. Mech. Eng.* 55: 129-160
- Liefooghe D, Fleury C 1990 Shape and mesh optimization using geometric and modeling methods. *AIAA J.* 28: 135-149
- Liu Y C, Elmaraghy H A 1992 Assessment of discretized errors and adaptive refinement with quadrilateral finite elements. *Int. J. Numer. Meth. Eng.* 33: 781-798
- Martha L F, Abel J F 1987 Tests of quadratic isoparametric solid element families. *Proceedings of NAFEMS International Conference 'Quality Assurance in Finite Element Analysis'*, Brighton
- Martinez R, Samartin A 1991 Two-dimensional mesh optimization in the finite element method. *Comput. Struc.* 40: 1169-1175
- McLeod R 1976 Node requirements for high-order approximation over curved finite elements. *J. Inst. Math. Appl.* 17: 249-254
- McLeod R 1978 Higher order transformation methods for curved finite elements. *J. Inst. Math. Appl.* 21: 419-428
- McLeod R, Mitchell A R 1972 The construction of basis functions for curved elements in the finite element method. *J. Inst. Math. Appl.* 10: 382-393
- McLeod R J Y, Mitchell A R 1975 The use of parabolic arcs in matching curved boundaries in the finite element method. *J. Inst. Math. Appl.* 16: 239-246
- McNeal R H, Harder R L 1985 A proposed standard set of problems to test finite element accuracy. *Finite Elements Anal. Design* 1:1
- McNiece G M, Marcal P V 1973 Optimization of finite element grids based on minimum potential energy. *Trans. ASME, J. Eng. Ind.* 95: 186-190
- Melosh R J 1963 Structural analysis of solids. *Am. Soc. Civil Eng.* ST4, 89: 205-223
- Melosh R J, Lobitz D W 1975 On numerical sufficiency test for monotonic convergence of finite element models. *AIAA J.* 13: 675-678
- Melosh R J, Marcal P V 1977 An energy basis for mesh refinement of structural continua. *Int. J. Numer. Meth. Eng.* 11: 1083-1091
- Mitchell A R 1979 Advantages of cubics for approximating element boundaries. *Comput. Maths. Appl.* 5: 321-327
- Mitchell A R, Phillips G, Wachspress E 1971 Forbidden shapes in the finite element method. *J. Inst. Math. Appl.* 8: 260-269
- Molinari G 1979 Grid iteration method for finite element grid optimization. *Finite element grid optimization* (ed.) M S Shephard, R H Gallagher (New York: ASME) PVP-38, pp. 1-14
- Newton R E 1973 Degeneration of brick-type isoparametric elements. *Int. J. Numer. Meth. Eng.* 7: 579-581

- Nicolas V T, Citipitioglu E 1977 A general isoparametric finite element program SDRC SUPERB. *Comput. Struc.* 17: 303–313
- Oddy A, Goldak M, McDill M, Bibby M 1988 A distortion metric for isoparametric finite elements. *Can. Soc. Mech. Eng.* 12: No. 4
- Okabe M 1981 Analytical integral formulae related to convex quadrilateral finite elements. *Comput. Meth. Appl. Mech. Eng.* 29: 201–218
- Parthasarathy V N, Kodiyalam S 1990 A constrained optimization approach to finite element mesh smoothing. *Finite Elements Anal. Design* 9: 309–320
- Peano A 1987 Inadmissible distortion of solid elements and patch test results. *Commun. Appl. Numer. Meth.* 3: 97–101
- Peano A, Pasini A 1982 A warning against misuse of quarter point finite elements. *Int. J. Numer. Meth. Eng.* 18: 314–320
- Rigby G L, McNeice G M 1972 A strain energy basis for studies of element stiffness matrices. *AIAA J.* 10: 1490–1493
- Robinson J 1976 A single element test. *Comput. Meth. Appl. Mech. Eng.* 7: 191–200
- Robinson J 1985 A new look at element shape parameters. *National Agency for Finite Element Methods and Standards*, Technical Report
- Robinson J 1987a Some new distortion measures for quadrilaterals. *Finite Elements Anal. Design* 3: 183–197
- Robinson J 1987b CRE method of element testing and the Jacobian shape parameters. *Eng. Comput.* 4:
- Robinson J 1988 Distortion measures for quadrilaterals with curved boundaries. *Finite Elements Anal. Design* 4: 115–131
- Robinson J 1990 Element shape sensitivity testing: the CRE-method and without a FEM system. *Finite Elements Anal. Design* 7: 73–84
- Robinson J 1991 Validity of aspect ratio sensitivity testing – An analytical investigation. *Finite Elements Anal. Design* 9: 125–132
- Salmon D C, Abel J F 1989 Assessing the effect of shape distortion in Q9 membrane elements. *Comput. Struc.* 33: 1183–1190
- Shephard M S 1979 Finite element grid optimization – A review. *Finite element Grid Optimization* (ed.) M S Shephard, R H Gallagher (New York: ASME) PV-38, pp. 1–14
- Shephard M S, Gallagher R H, Abel J F 1980 The synthesis of near-optimum finite element meshes with interactive computer graphics. *Int. J. Numer. Meth. Eng.* 15: 1012–1039
- Steinmueller G 1974 Restrictions in the application of automatic mesh generation schemes by isoparametric co-ordinates. *Int. J. Numer. Meth. Eng.* 8: 289–294
- Stephenson J W, Manohar R 1979 Forbidden nodes for curved finite elements. *Int. J. Numer. Meth. Eng.* 14: 1421–1427
- Stricklin W S, Ho, Richardson E Q, Aisler W E 1977 On isoparametric vs. linear strain triangular element. *Int. J. Numer. Meth. Eng.* 11: 1041–1043
- Turcke D J, McNiece G M 1974 Guidelines for selecting finite element grids based on an optimization study. *Comput. Struc.* 4: 499–519
- Ulbin M, Hellen T K 1989 Post-processing techniques for assessing element distortion errors. *Eng. Comput.* 6:
- Wachspress E L 1971 A rational basis for function approximation. *J. Inst. Math. Appl.* 8: 57–68
- Winslow A M 1967 Numerical solution of the quasilinear poisson equation in a non-uniform triangular mesh. *J. Comput. Phys.* 2: 149–172
- Woodford G, Mitchell A R, McLeod R 1978 An analysis of a cubic isoparametric transformation. *Int. J. Numer. Meth. Eng.* 12: 1587–1595
- Yon J H, Yang D Y 1990 A three-dimensional rigid-plastic finite element analysis of bevel gear forging by using a remeshing technique. *Int. J. Mech. Sci.* 32: 277–291
- Yang H T Y, Heinsteins M, Shih J M 1989 Adaptive 2D finite element simulation of metal forming processes. *Int. J. Numer. Meth. Eng.* 28: 1409–1428
- Zamal M 1973 A remark on the serendipity family. *Int. J. Numer. Meth. Eng.* 7: 98–100





## Magneto-visco-elastic surface waves in stressed conducting media

SAMAR CHANDRA DAS<sup>1</sup>, D P ACHARYA<sup>2</sup> and P R SENGUPTA<sup>3</sup>

<sup>1</sup>Indian Institute of Mechanics of Continua, 201, Manicktala Main Road, Suite No. 42, Calcutta 700 054, India

<sup>2</sup>Department of Mathematics, Mahadevananda College, Barrackpore, India

<sup>3</sup>Department of Mathematics, University of Kalyani, Kalyani 741 235, India

MS received 23 November 1992; revised 9 March 1994

**Abstract.** The present paper is concerned with magneto-visco-elastic surface waves in conducting media involving time rate of strain and stress of first order, the media being under an initial stress of hydrostatic tension or compression. The theory of magneto-visco-elastic surface waves in a conducting medium involving time rate of strain and stress of first order is derived under an initial stress. The above general theory is then employed to characterise Rayleigh, Love and Stoneley waves. Results obtained in the above cases reduce to well-known classical results when viscosity and magnetic field are absent.

**Keywords.** Magneto-visco-elastic; first order; surface waves; initial stress; hydrostatic tension or compression.

### 1. Introduction

Surface waves play an important role in the study of earthquakes, seismology, geophysics and geodynamics. The theory of surface waves has been widely investigated and developed by Rayleigh (1885), Voigt (1887), Stoneley (1924), Ewing *et al* (1957, pp. 257–259, 311), Hunter (1960, pp. 1–57), Bland (1960, pp. 30–75), Flugge (1967, pp. 3–21) and Jeffreys (1959, pp. 35–38). As the inner parts of the earth are under considerable stress from the weight of the matter resting on its surface, we may suppose that the initial equilibrium stress is approximately of hydrostatic nature. Moreover, the earth is placed in its own magnetic-field. Therefore, the investigation presented in this paper may be of importance when surface waves propagate under initial stress, magnetic fields and the viscous nature of the medium are involved. It is believed that the considered problem has not been so far investigated.

The interplay of an electromagnetic field with the motion of deformable solids has also been undertaken by many investigators (Knopoff 1955; Banos 1956; Chadwick 1957; Suhubi 1965; Yu & Tang 1966; De & Sengupta 1971). Yu & Tang (1966) thoroughly discussed the dilatational and rotational waves in a magneto-elastic

initially stressed conducting medium. De & Sengupta (1971, 1972) investigated magneto-elastic waves and disturbances in initially stressed conducting media. Acharya & Sengupta (1978) investigated the problem of magneto-thermo-elastic surface waves in initially stressed conducting media. More recently, the effect of viscosity on the elastic surface waves is receiving greater attention from many investigators (Das & Sengupta 1990a, 1990b, 1992; Roy & Sengupta 1983a, 1983b).

In the present paper the authors investigate magneto-visco-elastic surface waves in a conducting medium under hydrostatic stress (tension or compression) paying special attention to Rayleigh, Love and Stoneley waves. Dispersion relations are derived for Rayleigh and Love and some comments on Stoneley waves are also included.

## 2. Basic equations

The equations of motion for a perfectly conducting elastic solid under initial stress (hydrostatic tension or compression) in a uniform magnetic field are (Yu & Tang 1966)

$$\rho \frac{\partial^2 u_i}{\partial t^2} = -p_0 \frac{\partial^2 u_i}{\partial x_j \partial x_j} + \mu_e H_0 \left( \frac{\partial H_i}{\partial x_1} - \frac{\partial H_1}{\partial x_i} \right) + \frac{\partial \tau_{ij}}{\partial x_j}, \quad (1)$$

$$H_i = H_0 \left( \frac{\partial u_i}{\partial x_1} - \frac{\partial u_1}{\partial x_i} \right), \quad i, j = 1, 2, 3,$$

where  $p_0$  is the hydrostatic tension or compression (tension when  $p_0 < 0$  and compression when  $p_0 > 0$ ),  $\tau_{ij}$  is the stress tensor over the initial stress,  $u_i$  is the displacement vector with respect to coordinates  $x_1, x_2, x_3$  and time  $t$ ,  $\rho$  is the density of the material,  $H_0$  is the intensity of the uniform magnetic field parallel to  $x_1$ -axis,  $\mu_e$  is the magnetic permeability.

## 3. Formulation of the problem

Let  $M_1$  and  $M_2$  be two electrically conducting charge free isotropic, homogeneous, visco-elastic, semi-infinite solid media in welded contact under an initial hydrostatic

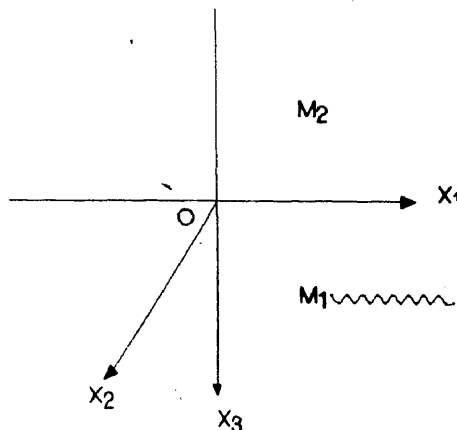


Figure 1. Interface geometry.

tension or compression permeated by uniform magnetic field (figure 1). We further assume that the medium still remains homogeneous and isotropic under the combined influence of magnetic field and initial stress. The magnetic properties of the conducting solids in the domains  $M_1$  and  $M_2$  are assumed to be sensibly the same. We consider a system of orthogonal Cartesian axes  $Ox_1x_2x_3$ , the origin  $O$  being on the interface, and  $Ox_3$  being normal to the interface (figure 1).

We consider the possibility of a type of wave travelling in the positive  $x_1$ -direction in such a manner that the disturbance is largely confined to the neighbourhood of the boundary and all the particles at any instant on any line parallel to  $x_2$ -axis have equal displacements. Due to the first assumption we assert that the wave is a surface wave and the second assumption concludes that all partial derivatives with respect to the coordinates  $x_2$  are zero. Then using the formulae  $u = \text{grad } \phi + \text{curl } \psi$ , the displacement components,  $u_1$  and  $u_3$  at any point may be expressed in the form

$$u_1 = \frac{\partial \phi}{\partial x_1} - \frac{\partial \psi}{\partial x_3}, \quad u_3 = \frac{\partial \phi}{\partial x_3} + \frac{\partial \psi}{\partial x_1}, \quad (2)$$

so that

$$\nabla^2 \phi = \Delta, \quad \nabla^2 \psi = \frac{\partial u_3}{\partial x_1} - \frac{\partial u_1}{\partial x_3}, \quad \nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_3^2}, \quad \Delta = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_3}{\partial x_3},$$

where  $\phi$  and  $\psi$  are functions of co-ordinates  $x_1, x_3$  and time  $t$ .

The first order stress-strain relation for an isotropic visco-elastic medium is (Voigt 1887)

$$\left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \tau_{ij} = \left( \lambda_1 + \lambda_2 \frac{\partial}{\partial t} \right) \Delta \delta_{ij} + 2 \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) e_{ij}, \quad (3)$$

where  $\eta_1, \lambda_1, \mu_1$  are elastic constants,  $\eta_2, \lambda_2$  and  $\mu_2$  are constants due to viscosity,  $e_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i})$  is the strain tensor and  $\delta_{ij}$  is the Kronecker symbol.

Using (3) in (1), the displacement equations of motion for a conducting first order visco-elastic medium under hydrostatic stress in an uniform magnetic field as

$$\begin{aligned} & \left[ (\lambda_1 + \mu_1) + (\lambda_2 + \mu_2) \frac{\partial}{\partial t} \right] \frac{\partial \Delta}{\partial x_1} + \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) \nabla^2 u_1 \\ & - \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) p_0 \nabla^2 u_1 = \rho \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \frac{\partial^2 u_1}{\partial t^2}, \\ & \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) \nabla^2 u_2 - p_0 \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \nabla^2 u_2 + K \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \frac{\partial^2 u_2}{\partial x_1^2} = \\ & \rho \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \frac{\partial^2 u_2}{\partial t^2}, \\ & \left[ (\lambda_1 + \mu_1) + (\lambda_2 + \mu_2) \frac{\partial}{\partial t} \right] \frac{\partial \Delta}{\partial x_3} + \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) \nabla^2 u_3 - p_0 \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \nabla^2 u_3 \\ & + K \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \left( \frac{\partial^2 u_3}{\partial x_1^2} - \frac{\partial^2 u_1}{\partial x_1 \partial x_3} \right) = \rho \left( \eta_1 + \eta_2 \frac{\partial}{\partial t} \right) \frac{\partial^2 u_3}{\partial t^2}, \end{aligned} \quad (4)$$

where  $K = \mu_e H_0^2$ . The above relations apply to both  $M_1$  and  $M_2$ .

Introducing (2) in (4) yields

$$\begin{aligned}\frac{\partial^2 \phi}{\partial t^2} &= \left[ \frac{(V_{1T}^2 + V_{2T}^2(\partial/\partial t))}{L} - \frac{p_0}{\rho} \right] \nabla^2 \phi + \frac{K}{\rho} \frac{\partial^2 \psi}{\partial x_1 \partial x_3}, \\ \frac{\partial^2 \psi}{\partial t^2} &= \left[ \frac{(V_{1S}^2 + V_{2S}^2(\partial/\partial t))}{L} - \frac{p_0}{\rho} \right] \nabla^2 \psi + \frac{K}{\rho} \frac{\partial^2 \psi}{\partial x_1^2}, \\ \frac{\partial^2 u_2}{\partial t^2} &= \left[ \frac{(V_{1S}^2 + V_{2S}^2(\partial/\partial t))}{L} - \frac{p_0}{\rho} \right] \nabla^2 u_2 + \frac{K}{\rho} \frac{\partial^2 u_2}{\partial x_1^2},\end{aligned}\quad (5)$$

where

$$\begin{aligned}V_{1T}^2 &= \frac{\lambda_1 + 2\mu_1}{\rho}, \quad V_{2T}^2 = \frac{\lambda_2 + 2\mu_2}{\rho}, \quad V_{1S}^2 = \frac{\mu_1}{\rho}, \quad V_{2S}^2 = \frac{\mu_2}{\rho}, \\ L &= \eta_1 + \eta_2 \frac{\partial}{\partial t}.\end{aligned}$$

Again, (5) apply to both  $M_1$  and  $M_2$ . In the sequel,  $M_2$  is identified by  $\rho', \eta'_1, \eta'_2, \lambda'_1, \lambda'_2, \mu'_1, \mu'_2$  for its properties.

### 3.1 Boundary conditions

To obtain the frequency equation we apply the following conditions of continuity across the interface

- (i) The components of displacement at the interface between  $M_1$  and  $M_2$  must be continuous.
- (ii) Stress components  $\tau_{31}, \tau_{32}, \tau_{33}$  are continuous across the interface. They are respectively given by

$$\begin{aligned}L\tau_{31} &= \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) \left( 2 \frac{\partial^2 \phi}{\partial x_1 \partial x_3} + \frac{\partial^2 \psi}{\partial x_1^2} - \frac{\partial^2 \psi}{\partial x_3^2} \right), \\ L\tau_{32} &= \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) \frac{\partial u_2}{\partial x_3}, \\ L\tau_{33} &= \left( \lambda_1 + \lambda_2 \frac{\partial}{\partial t} \right) \nabla^2 \phi + 2 \left( \mu_1 + \mu_2 \frac{\partial}{\partial t} \right) \left( \frac{\partial^2 \phi}{\partial x_3^2} + \frac{\partial^2 \psi}{\partial x_1 \partial x_3} \right).\end{aligned}\quad (6)$$

## 4. Harmonic solutions

To investigate equations (5), let us take harmonic solutions

$$(\phi, \psi, u_2) = [\hat{\phi}(x_3), \hat{\psi}(x_3), \hat{u}_2(x_3)] e^{i(\eta x_1 - \omega t)}, \quad (7)$$

for medium  $M_1$ . For medium  $M_2$ , the dashed functions  $\hat{\phi}', \hat{\psi}', \hat{u}_2'$  replace  $\hat{\phi}, \hat{\psi}, \hat{u}_2$ .

Introducing (7) in (5) we get

$$\frac{d^2 \hat{\phi}}{dx_3^2} - \left[ \eta^2 - \frac{\omega^2 \eta_K^*}{V_{KT}^2 - \eta_K^*(p_0/\rho)} \right] \hat{\phi} = \frac{-i\eta K \eta_K^*}{\rho(V_{KT}^2 - (p_0 \eta_K^*/\rho))} \frac{d\hat{\psi}}{dx_3},$$

$$\frac{d^2 \hat{\psi}}{dx_3^2} - \left[ \eta^2 - \frac{\omega^2 \eta_K^*}{V_{KS}^2 - \eta_K^* (p_0/\rho)} + \frac{K \eta^2 \eta_K^*}{\rho (V_{KS}^2 - \eta_K^* (p_0/\rho))} \right] \hat{\psi} = 0, \quad (8)$$

$$\frac{d^2 \hat{u}_2}{dx_3^2} - \left[ \eta^2 + \frac{((K/\rho) \eta^2 - \omega^2) \eta_K^*}{V_{KS}^2 - (p_0/\rho) \eta_K^*} \right] \hat{u}_2 = 0,$$

in which

$$\eta_K^* = (\eta_1 - i\omega\eta_2), \quad V_{KT}^2 = V_{1T}^2 - i\omega V_{2T}^2, \quad V_{KS}^2 = V_{1S}^2 - i\omega V_{2S}^2.$$

Similar relations for  $M_2$  can be obtained by using dashed variables  $\hat{\phi}', \hat{\psi}', \hat{u}_2', \eta_1', \eta_2', V_{1T}', V_{2T}', V_{1S}', V_{2S}', \eta_K^{*'}, V_{KT}'^2, V_{KS}'^2, \lambda_1', \mu_1', \lambda_2', \mu_2', \rho'$ . According to our assumption  $K$  is the same for both the media.

Clearly, (8) must have exponential solutions; and, in order that  $\phi, \psi$  and  $u_2$  describe surface waves, they must become vanishingly small as  $x_3$  tends to infinity. Thus for the medium  $M_1$  the solutions of (5) may be taken in the following form:

$$\begin{aligned} \phi &= [A \exp\{-x_3(\eta^2 - \zeta_1^2)^{1/2}\} + \beta \exp\{-x_3(\eta^2 - \zeta_2^2)^{1/2}\}] \exp\{i(\eta x_1 - \omega t)\}, \\ \psi &= B_1 \exp\{-x_3(\eta^2 - \zeta_2^2)^{1/2} + i(\eta x_1 - \omega t)\}, \\ u_2 &= C \exp\left\{-x_3 \left[ \eta^2 + \frac{(K/\rho) \eta^2 - \omega^2}{V_{KS}^2 - (p_0/\rho) \eta_K^*} \right]^{1/2} + i(\eta x_1 - \omega t)\right\}. \end{aligned} \quad (9)$$

For the medium  $M_2$

$$\begin{aligned} \phi' &= [A' \exp\{x_3(\eta'^2 - \zeta_1'^2)^{1/2}\} + B' \exp\{x_3(\eta'^2 - \zeta_2'^2)^{1/2}\}] \exp\{i(\eta x_1 - \omega t)\}, \\ \psi' &= B_1' \exp\{x_3(\eta'^2 - \zeta_2'^2)^{1/2} + i(\eta x_1 - \omega t)\}, \\ u_2' &= C' \exp\left\{x_3 \left[ \eta'^2 + \frac{((K/\rho') \eta'^2 - \omega^2) \eta_K^{*'}}{V_{KS}'^2 - (p_0/\rho') \eta_K^{*'}} \right]^{1/2} + i(\eta x_1 - \omega t)\right\}, \end{aligned} \quad (10)$$

where

$$\begin{aligned} \zeta_1^2 &= \frac{\omega^2 \eta_K^*}{V_{KT}^2 - \eta_K^* (p_0/\rho)}, & \zeta_1'^2 &= \frac{\omega^2 \eta_K^{*'}}{V_{KT}'^2 - \eta_K^{*'} (p_0/\rho')}, \\ \zeta_2^2 &= \frac{(\omega^2 - (K/\rho) \eta^2) \eta_K^*}{V_{KS}^2 - \eta_K^* (p_0/\rho)}, & \zeta_2'^2 &= \frac{(\omega^2 - (K/\rho') \eta'^2) \eta_K^{*'}}{V_{KS}'^2 - \eta_K^{*'} (p_0/\rho')}, \end{aligned} \quad (11)$$

and

$$B = \alpha_1 B_1, \quad B' = \alpha_1' B_1',$$

with

$$\alpha_1 = \frac{(i\eta K/\rho)(\eta^2 - \zeta_2^2)^{1/2}}{\omega^2 - \zeta_2^2 [(V_{KT}^2/\eta_K^*) - (p_0/\rho)]}, \quad \alpha_1' = \frac{(-i\eta K/\rho')(\eta'^2 - \zeta_2'^2)^{1/2}}{\omega^2 - \zeta_2'^2 [(V_{KT}'^2/\eta_K^{*'}) - (p_0/\rho')]}.$$

In evaluating quantities like  $(\eta^2 - \zeta^2)^{1/2}$ , the root with positive real part will be taken in each case.

Now applying the boundary conditions (i) and (ii) we get

$$A + (\alpha_1 - iQ_2)B_1 = A' + (\alpha_1' + iQ_2')B_1', \quad (12a)$$

$$C = C', \quad (12b)$$

$$iQ_1 A + (iQ_2 \alpha_1 + 1)B_1 = -iQ_1' A' - (i\alpha_1' Q_2' - 1)B_1', \quad (12c)$$

$$\rho \left( \frac{V_{KS}^2}{\eta_K^*} \right) [2iQ_1 A + \{(1 + Q_2^2) + 2iQ_2 \alpha_1\} B_1] \\ = \rho' \left( \frac{V_{KS}'^2}{\eta_K^{*'}} \right) [-2iQ_1' A' + \{-2iQ_2' \alpha_1' + (1 + Q_2'^2)\} B_1'], \quad (12d)$$

$$- \rho \left( \frac{V_{KS}^2}{\eta_K^*} \right) \left\{ \eta^2 + \frac{((K/\rho)\eta^2 - \omega^2)\eta_K^*}{V_{KS}^2 - (p_0/\rho)\eta_K^*} \right\}^{1/2} C \\ = \rho' \left( \frac{V_{KS}'^2}{\eta_K^{*'}} \right) \left\{ \eta'^2 + \frac{((K/\rho')\eta'^2 - \omega^2)\eta_K^{*'}}{V_{KS}'^2 - (p_0/\rho')\eta_K^{*'}} \right\}^{1/2} C', \quad (12e)$$

$$\left( \frac{\rho}{\eta_K^*} \right) [\{V_{KT}^2(Q_1^2 - 1) + 2V_{KS}^2\} A + \{V_{KT}^2(Q_2^2 - 1)\alpha_1 + 2V_{KS}^2(\alpha_1 - iQ_2)\} B_1] \\ = \left( \frac{\rho'}{\eta_K^{*'}} \right) [\{V_{KT}'^2(Q_1'^2 - 1) + 2V_{KS}'^2\} A' + \\ + \{V_{KT}'^2(Q_2'^2 - 1)\alpha_1' + 2V_{KS}'^2(\alpha_1' + iQ_2')\} B_1'] \quad (12f)$$

where

$$Q_k = \left( 1 - \frac{\zeta_K^2}{\eta^2} \right)^{1/2}, \quad Q_k' = \left( 1 - \frac{\zeta_K'^2}{\eta'^2} \right)^{1/2}, \quad k = 1, 2.$$

From (12b) and (12e) we get

$$C = C' = 0.$$

Thus we conclude that there is no propagation of the displacement  $u_2$ . The wave velocity equation is, therefore, obtained from (12a), (12c), (12d), (12f) by the elimination of the constants  $A$ ,  $B_1$ ,  $A'$ ,  $B_1'$ , in the determinant form as

$$|M_{ij}| = 0, \quad (i, j = 1, 2, 3, 4), \quad (13)$$

where

$$M_{11} = 1, \quad M_{12} = (\alpha_1 - iQ_2), \quad M_{13} = -1, \quad M_{14} = -(\alpha_1' + iQ_2'), \\ M_{21} = iQ_1, \quad M_{22} = (iQ_2 \alpha_1 + 1), \quad M_{23} = iQ_1', \quad M_{24} = (i\alpha_1' Q_2' - 1), \\ M_{31} = \rho(V_{KS}^2/\eta_K^*)2iQ_1, \quad M_{32} = \rho(V_{KS}^2/\eta_K^*)\{(1 + Q_2^2) + 2iQ_2 \alpha_1\}, \\ M_{33} = \rho'(V_{KS}'^2/\eta_K^{*'})2iQ_1', \quad M_{34} = -\rho'(V_{KS}'^2/\eta_K^{*'})\{-2iQ_2' \alpha_1' + (1 + Q_2'^2)\}, \\ M_{41} = (\rho/\eta_K^*)\{V_{KT}^2(Q_1^2 - 1) + 2V_{KS}^2\}, \\ M_{42} = (\rho/\eta_K^*)\{V_{KT}^2(Q_2^2 - 1)\alpha_1 + 2V_{KS}^2(\alpha_1 - iQ_2)\}, \\ M_{43} = -(\rho'/\eta_K^{*'})\{V_{KT}'^2(Q_1'^2 - 1) + 2V_{KS}'^2\}, \\ M_{44} = -(\rho'/\eta_K^{*'})\{V_{KT}'^2(Q_2'^2 - 1)\alpha_1' + 2V_{KS}'^2(\alpha_1' + iQ_2')\}.$$

From (13) we obtain the wave velocity in the common boundary under consideration in the presence of magnetic field, initial stress in the nature of hydrostatic tension or compression, viscosity where viscosity is of first order including strain rate and stress rate.

## 5. Particular cases

### 5.1 Rayleigh waves

In the particular case of Rayleigh waves the interface becomes a free surface and  $M_2$  is treated as vacuum.

Hence in view of (12d) and (12f) we get

$$2iQ_1 A + \{(1 + Q_2^2) + 2iQ_2 \alpha_1\} B_1 = 0, \quad (14)$$

$$\{V_{KT}^2(Q_1^2 - 1) + 2V_{KS}^2\} A + \{V_{KT}^2(Q_2^2 - 1)\alpha_1 + 2V_{KS}^2(\alpha_1 - iQ_2)\} B_1 = 0. \quad (15)$$

For the indispensable constants  $A, B_1$  from (14) and (15) to assume non zero values we have

$$|M'_{ij}| = 0 \quad (i, j = 1, 2), \quad (16)$$

where

$$M'_{11} = 2iQ_1, \quad M'_{12} = (1 + Q_2^2) + 2iQ_2 \alpha_1,$$

$$M'_{21} = V_{KT}^2(Q_1^2 - 1) + 2V_{KS}^2, \quad M'_{22} = V_{KT}^2(Q_2^2 - 1)\alpha_1 + 2V_{KS}^2(\alpha_1 - iQ_2).$$

Equation (16) represents the magneto-visco-elastic Rayleigh wave velocity equation under the initial stress in the nature of hydrostatic tension or compression in a medium, including strain rate and stress rate.

In the absence of viscous effects, (16) reduces to

$$\left| \frac{2iP_1}{V_{1T}^2(P_1^2 - 1) + 2V_{1S}^2} \quad \frac{(1 + P_2^2) + 2in_1 P_2}{V_{1T}^2(P_2^2 - 1)n_1 + 2V_{1S}^2(n_1 - iP_2)} \right| = 0, \quad (17)$$

where

$$P_1^2 = 1 - \frac{\omega^2}{(V_{1T}^2 - (p_0/\rho))\eta^2}, \quad P_2^2 = 1 - \frac{\omega^2 - K\eta^2/\rho}{(V_{1S}^2 - (p_0/\rho))\eta^2},$$

$$n_1 = \frac{iK[1 - (\omega^2/\eta^2 V_{1S}^2) + (K/\rho V_{1S}^2)]^{1/2}}{\rho V_{1T}^2[(\omega^2/\eta^2 V_{1T}^2) - (\omega^2/\eta^2 V_{1S}^2) + (KV_{1T}^2/\rho V_{1S}^2)]}.$$

Equation (17) represents the magneto-elastic Rayleigh-wave velocity equation under the initial stress of hydrostatic tension or compression. This agrees with the result obtained by Acharya & Sengupta (1978).

Moreover, in the absence of the magnetic field and initial stress ( $p_0 = 0, K = 0$ ) we get from (17) the Rayleigh wave velocity equation for the elastic medium as

$$4[1 - (c^2/V_{1T}^2)]^{1/2}[1 - (c^2/V_{1S}^2)]^{1/2} = (2 - (c^2/V_{1S}^2))^2, \quad (18)$$

where

$$c^2 = \omega^2/\eta^2.$$

Equation (18) is in complete agreement with the classical result of Rayleigh.

### 5.2 Love waves

We know that for such types of waves  $u_2$  is the only component of displacement vector  $\mathbf{u}$  to play the role. Let us consider that the medium  $M_2$  is bounded by two



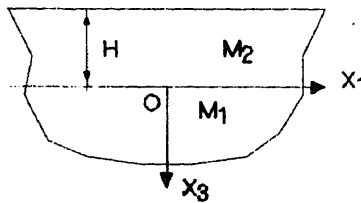


Figure 2. Love wave formulation.

horizontal plane surfaces at a finite distance  $H$  apart, the upper plane surface being free while the medium  $M_1$  extends to an infinitely great distance (figure 2).

The notable fact here is that the displacement in  $M_2$  may no longer diminish with distance from the boundary between  $M_1$  and  $M_2$  so that for the medium  $M_2$  we preserve the full solution as

$$u'_2 = \left[ C'_1 \exp \left\{ x_3 \left( \eta^2 + \frac{((K/\rho')\eta^2 - \omega^2)\eta_K^{*'}}{V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}} \right)^{1/2} \right\} + C'_2 \exp \left\{ -x_3 \left( \eta^2 + \frac{((K\eta^2/\rho') - \omega^2)\eta_K^{*'}}{(V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'})} \right)^{1/2} \right\} \right] \times \exp(i(\eta x_1 - \omega t)), \quad (19)$$

where the restriction that the real part of  $\{\eta^2 + ((K/\rho')\eta^2 - \omega^2)\eta_K^{*'}/[V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}]\}^{1/2}$  be positive is not required.

For the present case the boundary conditions are

- (i)  $u_2$  and  $\tau_{32}$  are continuous at  $x_3 = 0$ ,
- (ii)  $\tau'_{32} = 0$  at  $x_3 = -H$ .

Employing the boundary conditions (i) and (ii) we obtain

$$C = C'_1 + C'_2 \quad (20a)$$

$$-\rho \left( \frac{V_{KS}^2}{\eta_K^*} \right) \left[ \eta^2 + \frac{((K\eta^2/\rho) - \omega^2)\eta_K^*}{V_{KS}^2 - (p_0/\rho)\eta_K^*} \right]^{1/2} C = \rho' \left( \frac{V_{KS}^{\prime 2}}{\eta_K^{*'}} \right) \times \left[ \eta^2 + \frac{((K\eta^2/\rho') - \omega^2)\eta_K^{*'}}{V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}} \right]^{1/2} (C'_1 - C'_2), \quad (20b)$$

$$C'_1 e^{-H} \left[ \eta^2 + \frac{((K/\rho')\eta^2 - \omega^2)\eta_K^{*'}}{V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}} \right]^{1/2} - C'_2 e^{-H} \left[ \eta^2 + \frac{((K\eta^2/\rho') - \omega^2)\eta_K^{*'}}{V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}} \right]^{1/2} = 0. \quad (20c)$$

Eliminating  $C$ ,  $C'_1$ ,  $C'_2$  from equations (20a)–(20c) we get

$$\rho \left( \frac{V_{KS}^2}{\eta_K^*} \right) \left[ 1 - \frac{(c^2 - (K/\rho))\eta_K^*}{V_{KS}^2 - (p_0/\rho)\eta_K^*} \right]^{1/2} - \rho' \left( \frac{V_{KS}^{\prime 2}}{\eta_K^{*'}} \right) \left[ \frac{(c^2 - (K/\rho'))\eta_K^{*'}}{V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}} - 1 \right]^{1/2} \times \tan \left\{ \eta H \left[ \frac{(c^2 - (K/\rho'))\eta_K^{*'}}{V_{KS}^{\prime 2} - (p_0/\rho')\eta_K^{*'}} - 1 \right]^{1/2} \right\} = 0, \quad (21)$$

where  $c = \omega/\eta$ .

Equation (21) represents the required wave velocity equations for Love waves in a magneto-visco-elastic solid medium including strain and stress rates of first order under an initial hydrostatic tension or compression. It is seen from (21) that Love waves depend upon viscous field as well as magnetic field and also on the initial stress in the nature of hydrostatic tension or compression.

### 5.3 Stoneley waves

In the classical theory, Stoneley waves are the generalised form of Rayleigh waves propagating in the vicinity of interface of two semi-infinite media  $M_1$  and  $M_2$ . Hence in our general case Stoneley waves propagating along the common boundary of  $M_1$  and  $M_2$  are determined by the roots of the wave velocity equation (13). In the absence of magnetic field, initial stress, viscosity and strain-stress rates this equation of course reduces to the classical result obtained by Stoneley.

We are grateful to the referees for their valuable comments and suggestions towards improvement of the paper and its appearance in the present form.

### References

- Acharya D P, Sengupta P R 1978 Magneto-thermo-elastic surface waves in initially stressed conducting media. *Acta Geophys. Polon.* 26(4):
- Banos A 1956 Normal modes characterising magneto-elastic plane waves. *J. Phys. Rev.* 104: 300–305
- Bland D R 1960 *The theory of linear visco-elasticity* (London: Pergamon) (This monograph on the subject contains many cases of stress analysis)
- Chadwick P 1957 Elastic wave propagation in a magnetic field. *IX Congress Int. Mech. Appl.* 7: 143–158
- Das T K, Sengupta P R 1990a Surface waves in general viscoelastic media of higher order. *Indian. J. Pure Appl. Math.* 21(7): 661–675
- Das T K, Sengupta P R 1990b Surface waves in thermo-visco-elastic media considering time rate of stress and strain of higher order. *Gerlands Beitr. Geophys. Leipzig* 99: 337–448
- Das T K, Sengupta P R 1992 Effect of gravity on visco-elastic surface waves in solids involving time rate of strain and stress of first order. *Sādhana* 17: 315–323
- De S N, Sengupta P R 1971 Surface waves in magneto-elastic initially stressed conducting media. *Pure Appl. Geophys.* 88: 44–52
- De S N, Sengupta P R 1972 Magneto-elastic waves and disturbances in initially stressed conducting media. *Pure Appl. Geophys.* 93: 41–54
- Ewing W M, Jardetzky W S, Press F 1957 *Elastic waves in layered media* (London: McGraw-Hill)
- Flügge W 1967 *Visco-elasticity* (London: Blaisdell)
- Hunter S C 1960 *Visco-elastic waves. Progress in solid mechanics* (eds) I N Snedon, R Hill (Amsterdam, New York: North Interscience)
- Jeffreys H 1959 *The earth* 4th edn (Cambridge: University Press)
- Knopoff L 1955 The interaction between elastic wave motions and a magnetic field in electric conductors. *J. Geophys. Res.* 60: 441–456
- Rayleigh Lord 1885 On waves propagated along the plane surface of an elastic solid. *Proc. London Math. Soc.* 17: 4–11
- Roy S K, Sengupta P R 1983a Rotatory vibration of a sphere of general visco-elastic solid. *Gerlands Beitr. Geophys. Leipzig* 92: 70–76
- Roy S K, Sengupta P R 1983b Radial vibration of a sphere of general visco-elastic solid. *Gerlands Beitr. Geophys. Leipzig* 92: 435–442

- Stoneley R 1924 Elastic waves at the surface of separation of two solids. *Proc. R. Soc., London* 106: 416-428
- Suhubi E S 1965 Small torsional oscillations of a circular cylinder with finite electric conductivity in a constant axial magnetic field. *Int. J. Eng. Sci.* 2: 441-459
- Voigt W 1887 Theoretische Studien über die Elasticitäts Verhältnisse der Krystalle. *Abh. Ges. Wiss. Göttingen* 34
- Yu C P, Tang S 1966 Magneto-elastic waves in initially stressed conductors. *Z. Angew. Math. Phys.* 17: 766

# Advances in Aerospace Engineering

## Foreword

This issue of *Sādhanā* contains seven papers selected from those presented at the International Symposium held in celebration of the Golden Jubilee of the Department of Aerospace Engineering, Indian Institute of Science, during 14–16 December 1992. The papers cover different aspects of aerospace science and technology, and concern the most recent developments in the state of the art.

The first paper, by S Selvarajan and V Vasanta Ram, discusses the stability of flow in a channel whose walls are excited by wave-like motion. It is shown theoretically that there is a wavelength-amplitude band in which the flow is stabilized, a finding that suggests possibilities of managing flow transition through wall excitation.

A Das reviews the numerical methods being developed at DLR, Braunschweig to handle flow problems that are complex either because of physics (involving say vortices, shock waves, shock/boundary-layer interactions etc.) or because of geometry (complete aerospace vehicle configurations, for example). A large number of complex flow fields are studied, including delta wings, canard configurations, propellers and rotors etc.

V J Modi and T Yokomizo report on their latest studies of moving-surface boundary layer control. Rotating cylinders at the leading and/or trailing edge of a Joukowski aerofoil are shown to yield substantial increases in lift and delays in stall. Beneficial effects are found on bluff bodies as well, with a reduction in drag and delay of separation.

Inderjit Chopra presents an assessment of recent trends in the design and analysis of helicopter rotor systems, in particular hingeless, bearing-less, composite, circulation-control, tilt and other advanced geometry rotors.

K Appa and J Argyris describe a multidisciplinary package that combines computational fluid dynamics, structural flexibility and flight control analysis to provide a new methodology for prediction of flight loads.

A R Upadhyaya and K Panda review recent progress in dynamics and aeroelasticity, with special emphasis on aeroelastic tailoring, structural optimisation and aeroservoelasticity.

A Nosier, R K Kapania and J N Reddy present a critical comparison of several theories to obtain the low-velocity impact response of laminated plates. The layer-wise plate theory is found to explain several phenomena including matrix cracking, fibre breakage, debonding etc.

We hope that these seven papers, all at the frontiers of research in their respective topics, will be of wide interest to engineers and scientists, especially in the aerospace field.

[REDACTED]

## Dynamical characteristics of wave-excited channel flow

S SELVARAJAN<sup>1</sup> and V VASANTA RAM<sup>2</sup>

<sup>1</sup>National Aeronautical Laboratory, Vimanapura Road, Bangalore 560017, India

<sup>2</sup>Institut für Thermo- und Fluidodynamik, Ruhr-Universität, Bochum, Germany

**Abstract.** This paper is part of a study on the receptivity characteristics of the shear flow in a channel whose walls are subjected to a wave-like excitation. The small amplitude forced wavy wall motion is characterised by a wave number vector  $\lambda_1, \lambda_2$  and a frequency  $\omega_g$ . The basic flow in the problem is a superposition of the Poiseuille flow and a periodic component that corresponds to the wave excitation of the wall. The aim of the study is to examine the susceptibility of this flow to transition. The problem is approached through studying the stability characteristics of the basic flow with respect to small disturbances. The theoretical framework for this purpose is Floquet theory. The solution procedure for solving the eigenvalue problem is the spectral collocation method. Preliminary results showing the influence of the amplitude and the wave number of the wall excitation on the stability boundary of the flow are presented.

**Keywords.** Dynamical characteristics; wave-excited channel flow; forced wavy wall motion; stability characteristics; spectral collocation method.

### 1. Introduction

In the broad area of study under the title “the stability of fluid motions”, a subject that has attracted increased attention in recent years is the response to further disturbances of the shear flow that is excited by waves travelling in the plane of the walls. Besides the inherent fundamental interest that disturbance-propagation problems hold as such in fluid mechanics, the possibility of influencing transition and thus of managing turbulence through wall excitation has been recognised for some time now and this has lent additional impetus to research on this subject. It is therefore not surprising to find this subject occupying a prominent position in papers at conferences on turbulence control and management (see e.g. Liepmann & Narasimha 1988), and at special sessions at larger conferences. Of the latter we mention here only the first European Fluid Mechanics Conference, 1991 (Cambridge, UK), The SIAM Conference on Applications of Dynamical Systems, 1992 (Salt Lake City), and the Fourth European Turbulence Conference 1992 (Delft, Netherlands).

In studies of disturbance propagation in these flows the points of main interest centre around the effects of the wall-wave excitation parameters, viz. its frequency,

wave number (vector) and amplitude, on the growth (or decay) characteristics of the further disturbances. At moderate Reynolds numbers one is particularly interested in the shift in the surface of neutral stability due to the wave excitation at the wall since this is an indication of the susceptibility of the flow to transition to turbulence. When the Reynolds numbers are much larger, the propagation of disturbances has to be studied in a flow that is already turbulent and this belongs to the realm of "management" of turbulence through the wave excitation of the walls. The latter context requires for its understanding a much deeper insight into the turbulence mechanism in the flow than is available today. It poses a problem of formidable diversity and complexity. As against this, the former has the advantage of being more tractable, yet retaining some, if not all, physical elements that characterise a turbulent flow. It is therefore better suited for study at the present stage and therefore defines the scope of this work. The disturbance propagation problem in the laminar flow with wave excited walls, despite its conceptual simplicity, is a difficult one, not only because it involves a larger number of parameters than the classical problem with unexcited walls, but also due to additional physical phenomenon that are encountered due to the wave-like component present in the basic flow. These call for a modification of the mathematical tools used to handle these problems.

The subject of the present paper is the propagation of small amplitude disturbances in the channel flow whose walls undergo a travelling wave type of deformation of one wave number and frequency only. We also regard the amplitude of the wave excitation at the wall to be sufficiently small, so that the basic flow in which disturbance propagation is to be studied comprises only one wave component superposed upon the classical fully developed flow with a parabolic velocity profile in a channel with rigid walls. The number of parameters involved in this problem is much larger than in the classical problem and its thorough investigation calls for a more extensive study than is possible in the relatively brief time span of around two years over which this work has been in progress. Even with the limited scope set, the study is too extensive to be accommodated within this paper, so we restrict ourselves for the present primarily to outlining our approach and presenting a selection of preliminary results that we have obtained so far. The reader interested in further details is referred to Selvarajan & Vasanta Ram (1991).

## 2. The basic flow

The equations of motion for an incompressible fluid are, in the usual notation,

$$\partial u_l / \partial x_l = 0, \quad l = 1, 2, 3 \quad (1a)$$

$$\frac{\partial u_l}{\partial t} + u_m \frac{\partial u_l}{\partial x_m} = - \frac{\partial p}{\partial x_l} + \frac{1}{\text{Re}} \frac{\partial^2 u_l}{\partial x_m \partial x_m}, \quad l, m = 1, 2, 3, \quad (1b)$$

where the lengths are referred to the mean semi-channel height  $H$ , the velocity components to the mean centre line velocity  $U_0$ , the time to  $H/U_0$ , the pressure to  $\rho U_0^2$  and the Reynolds number  $\text{Re}$  is based on  $H$  and  $U_0$ .

We specify the wall motion through the following expression (see figure 1),

$$y_w(t, x_1, x_3) = \pm 1 \mp \varepsilon_w \text{Real}(\exp[i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)]), \quad (2)$$

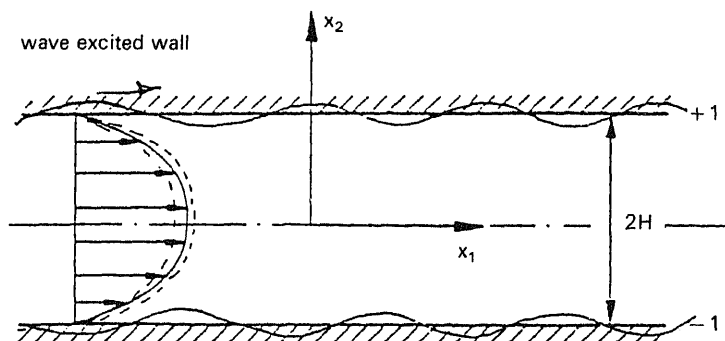


Figure 1. Flow configuration.

which represents a wave travelling in the wall plane around the position  $x_2 = \pm 1$  with amplitude  $\varepsilon_w$ , wave number vector  $(\lambda_1, \lambda_3)$  and frequency  $\omega_g$ . In the above form the waves on the upper and lower walls are in phase so that at any instant of time the local channel width expands and contracts around the value  $2H$ . It is relatively straightforward in principle to extend the study to permit the waves on the two walls to be out of phase with each other. This might be important for the behaviour of further disturbances and hence from the point of view of transition control. However, an additional parameter, viz. the phase difference, would then enter the picture. In the present state of such studies it is desirable to keep the number of additional parameters within limits and therefore we restrict ourselves in this work to walls excited in phase as in (2).

For small amplitudes of the wave excitation at the wall, the solution for the flow quantities may be sought as a perturbation from the parabolic velocity profile of the fully developed channel flow. To a linear approximation in  $\varepsilon_w$  we may then write for the flow quantities:

$$u_l = \delta_{1l}(1 - x_2^2) + \varepsilon_w \tilde{u}_l + O(\varepsilon_w^2), \quad l = 1, 2, 3, \quad (3a)$$

$$p = -(2x_1/\text{Re}) + \varepsilon_w \tilde{p} + O(\varepsilon_w^2). \quad (3b)$$

Transferring the boundary conditions in (2) to the mean position of the wall by standard methods (see, e.g. Van Dyke 1975) we get the boundary conditions for the perturbation  $\tilde{u}_l$  as follows:

$$\tilde{u}_1(\pm 1) = -2 \text{Real}(\exp[i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)]), \quad (4a)$$

$$\tilde{u}_2(\pm 1) = \text{Real}(\pm i\omega_g \exp[i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)]), \quad (4b)$$

$$\tilde{u}_3(\pm 1) = 0. \quad (4c)$$

The form of the solution for  $\tilde{u}_l$  may then be written as

$$\tilde{u}_l = \frac{1}{2}[\hat{u}_l(x_2)\exp[i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)] + \hat{u}_l^*(x_2)\exp[-i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)]], \quad l = 1, 2, 3, \quad (5a)$$

$$\tilde{p} = \frac{1}{2}[\hat{p}(x_2)\exp[i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)] + \hat{p}^*(x_2)\exp[-i(\lambda_1 x_1 + \lambda_3 x_3 - \omega_g t)]] \quad (5b)$$

where the superscript \* denotes the complex conjugate.



The complex amplitude functions for the velocity  $u_1$  and pressure  $\hat{p}$  in (5) then obey the following linearised equations of motion:

$$i\lambda_1 \hat{u}_1 + (d\hat{u}_2/dx_2) + i\lambda_3 \hat{u}_3 = 0, \quad (6a)$$

$$i[-\omega_g + (1 - x_2^2)\lambda_1]\hat{u}_1 - 2x_2\hat{u}_2 = -i\lambda_1 \hat{p} + (1/\text{Re})[(-\lambda_1^2 - \lambda_3^2) + (d^2/dx_2^2)]\hat{u}_1, \quad (6b)$$

$$i[-\omega_g + (1 - x_2^2)\lambda_1]\hat{u}_2 = -(d\hat{p}/dx_2) + (1/\text{Re})[(-\lambda_1^2 - \lambda_3^2) + (d^2/dx_2^2)]\hat{u}_2, \quad (6c)$$

$$i[-\omega_g + (1 - x_2^2)\lambda_1]\hat{u}_3 = -i\lambda_3 \hat{p} + (1/\text{Re})[(-\lambda_1^2 - \lambda_3^2) + (d^2/dx_2^2)]\hat{u}_3. \quad (6d)$$

The above equations (6a-d) are the linearised equations of motion from which the well-known Orr-Sommerfeld equations for investigations of the more classical studies on fluid flow stability are derived. The difference from the classical problem here is that in the present case they satisfy inhomogeneous boundary conditions due to wall excitation. These are:

$$\hat{u}_1(\pm 1) = -2; \quad \hat{u}_2(\pm 1) = \pm i\omega_g; \quad \hat{u}_3(\pm 1) = 0. \quad (7)$$

The set of equations (6) with the boundary conditions (7) have been solved numerically by two methods. One of the methods uses superposition coupled with an orthonormalization procedure with a variable-step Runge-Kutta-Fehlberg integration scheme (Scott & Watts 1977), and the other is the pseudo-spectral collocation method described in Gottlieb *et al* (1984, pp. 1-54). The solution procedures are outlined in Selvarajan & Vasanta Ram (1991). Plots of the velocity components for a set of parameters are shown in figures 2-7. The frequency  $\omega_g$  prescribed here is the same as the frequency of the neutrally stable Tollmien-Schlichting wave at the chosen wave number  $(\lambda_1, \lambda_3)$ .

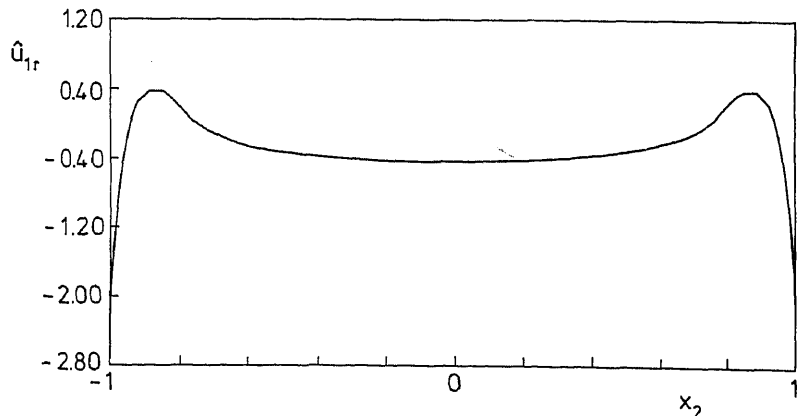
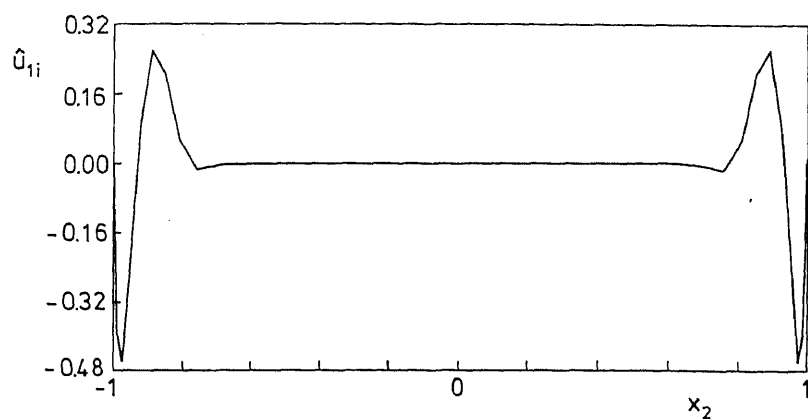
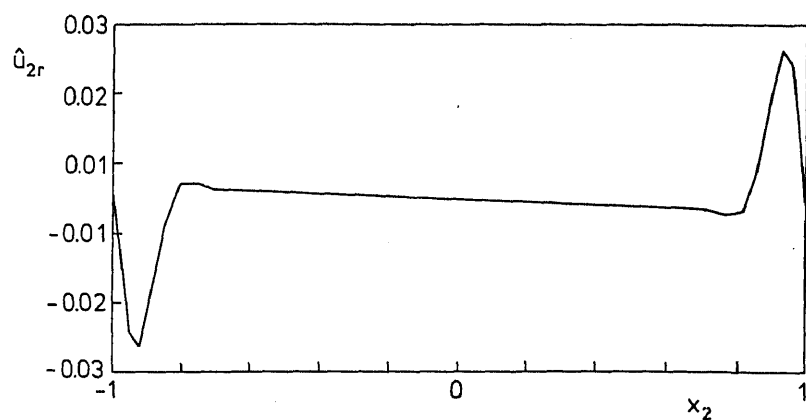


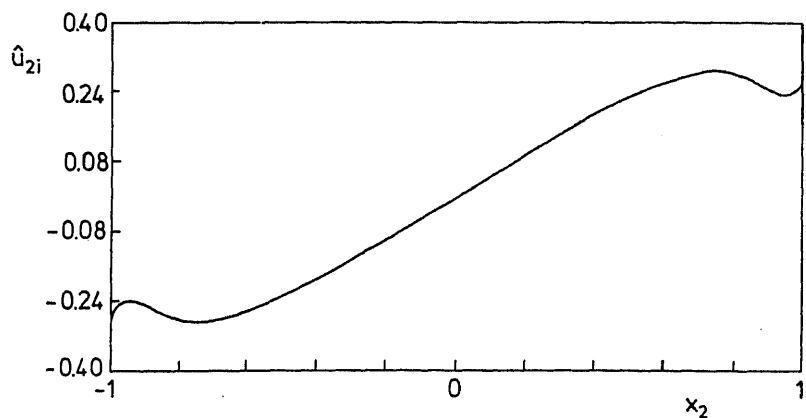
Figure 2. Variation of  $\hat{u}_{1,r}$ . Basic flow parameters:  $\lambda_1 = 1.04$ ;  $\lambda_3 = 0.29$ ;  $\omega_g = 0.27$ .



**Figure 3.** Variation of  $\hat{u}_{1i}$ . Basic flow parameters:  $\lambda_1 = 1.04$ ;  $\lambda_3 = 0.29$ ;  $\omega_g = 0.27$ .



**Figure 4.** Variation of  $\hat{u}_{2r}$ . Basic flow parameters:  $\lambda_1 = 1.04$ ;  $\lambda_3 = 0.29$ ;  $\omega_g = 0.27$ .



**Figure 5.** Variation of  $\hat{u}_{2i}$ . Basic flow parameters:  $\lambda_1 = 1.04$ ;  $\lambda_3 = 0.29$ ;  $\omega_g = 0.27$ .

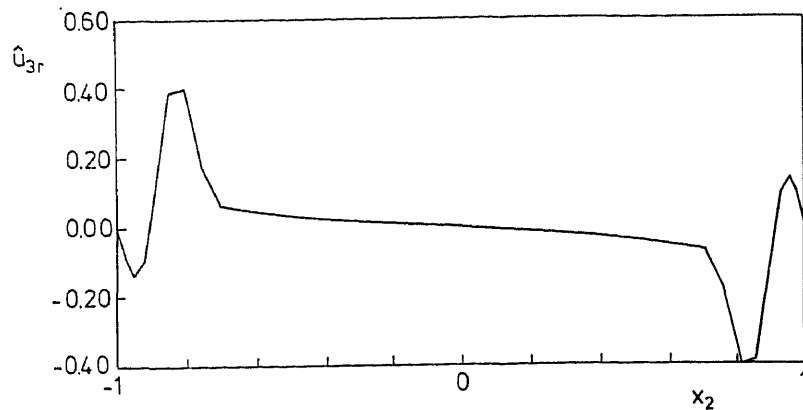


Figure 6. Variation of  $\hat{u}_{3r}$ . Basic flow parameters:  $\lambda_1 = 1.04$ ;  $\lambda_3 = 0.29$ ;  $\omega_g = 0.27$ .

### 3. The propagation of disturbances in the basic flow

The basic flow in which we study disturbance propagation in this work is defined through (3) above. Since this contains components that are periodic in time and in two space variables, the disturbance propagation problem has to be approached through the Floquet theory. For small amplitude, further disturbances of the order  $O(\varepsilon_s)$  over the basic flow, with  $\varepsilon_s \ll \omega$ , we may linearise the equations of motion to obtain the governing equations for the disturbance.

Writing the velocity and pressure as

$$u_l = \delta_{11}(1 - x_2^2) + \varepsilon_w \tilde{u}_l + \varepsilon_s u_{sl}, \quad l = 1, 2, 3, \quad (8a)$$

$$p = -(2x_1/\text{Re}) + \varepsilon_w \tilde{p} + \varepsilon_s p_s, \quad (8b)$$

the linearised equations of motion for the further disturbance are

$$\partial u_{sl} / \partial x_1 = 0, \quad (9a)$$

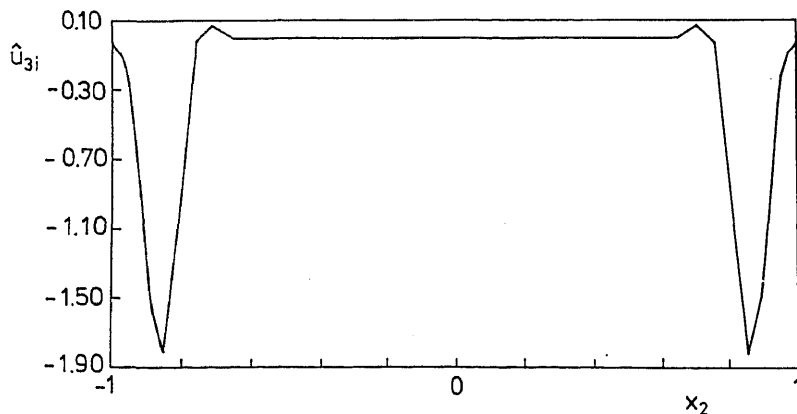


Figure 7. Variation of  $\hat{u}_{3i}$ . Basic flow parameters:  $\lambda_1 = 1.04$ ;  $\lambda_3 = 0.29$ ;  $\omega_g = 0.27$ .

$$\frac{\partial u_{s1}}{\partial t} + (1 - x_2^2) \frac{\partial u_{s1}}{\partial x_1} - 2x_2 u_{s2} + \frac{\partial p_s}{\partial x_1} + \varepsilon_w \left[ \tilde{u}_m \frac{\partial u_{s1}}{\partial x_m} + u_{sm} \frac{\partial \tilde{u}_1}{\partial x_m} \right] + \frac{1}{\text{Re}} \frac{\partial^2 u_{s1}}{\partial x_m \partial x_m} = 0, \quad (9b)$$

$$\frac{\partial u_{s2}}{\partial t} + (1 - x_2^2) \frac{\partial u_{s2}}{\partial x_1} + \frac{\partial p_s}{\partial x_2} + \varepsilon_w \left[ \tilde{u}_m \frac{\partial u_{s2}}{\partial x_m} + u_{sm} \frac{\partial \tilde{u}_2}{\partial x_m} \right] + \frac{1}{\text{Re}} \frac{\partial^2 u_{s2}}{\partial x_m \partial x_m} = 0, \quad (9c)$$

$$\frac{\partial u_{s3}}{\partial t} + (1 - x_2^2) \frac{\partial u_{s3}}{\partial x_1} + \frac{\partial p_s}{\partial x_3} + \varepsilon_w \left[ \tilde{u}_m \frac{\partial u_{s3}}{\partial x_m} + u_{sm} \frac{\partial \tilde{u}_3}{\partial x_m} \right] + \frac{1}{\text{Re}} \frac{\partial^2 u_{s3}}{\partial x_m \partial x_m} = 0. \quad (9d)$$

Our interest lies in the stability characteristics of the wave excited basic flow due to further disturbance, so homogeneous boundary conditions for  $u_{sl}$  are appropriate. It is consistent with the approximations to impose these conditions at the mean position of the channel walls  $x_2 = \pm 1$ .

#### 4. Form of solution for the eigenvalue problem (Floquet-Ansatz)

The outstanding feature of the disturbance equation (9) that demarcates it from more classical stability problems is the presence of periodic terms in  $t$ ,  $x_1$  and  $x_3$  through  $\tilde{u}_l$ . It is known (see e.g. Bender & Orszag 1984) that, although the parameter  $\varepsilon_w$  may be small, the structure of the solution is drastically changed by the periodic terms. The form of the solution for (9) that transforms the partial differential equation into an ordinary differential equation may be written as follows:

$$u_{sl} = \sum_{n=-\infty}^{n=+\infty} A_{sl}^{(n)}(x_2) \exp[i\varphi_s^{(n)}], \quad l = 1, 2, 3, \quad (10a)$$

$$p_s = \sum_{n=-\infty}^{n=+\infty} A_{sp}^{(n)}(x_2) \exp[i\varphi_s^{(n)}], \quad (10b)$$

where

$$\varphi_s^{(n)} = (\alpha_s + n\lambda_1)x_1 + (\beta_s + n\lambda_3)x_3 - (\omega_s + n\omega_g)t. \quad (10c)$$

The expression in (10) represents a superposition of waves of wave number  $(\alpha_s + n\lambda_1)$ ,  $(\beta_s + n\lambda_3)$  and frequency  $(\omega_s + n\omega_g)$ . In the temporal problem  $\alpha_s$  and  $\beta_s$  are regarded real. Solution of the problem then yields the complex frequency  $\omega_s$  for which (9) admits nontrivial solutions. The real part of  $\omega_s$  then indicates the growth or decay of the further disturbance characterised by  $(\alpha_s, \beta_s)$ . We write the dispersion relation of the problem symbolically in the following form

$$F(\omega_s, \alpha_s, \beta_s, \text{Re}, \varepsilon_w, \lambda_1, \lambda_3, \omega_g) = 0, \quad (11)$$

that shows the additional parameters that enter the problem due to the wave excitation of the wall. These are  $\varepsilon_w$ ,  $\lambda_1$ ,  $\lambda_3$  and  $\omega_g$ .

The equations governing the eigenfunctions  $A_{sl}^{(n)}(x_2)$  and  $A_{sp}^{(n)}(x_2)$  are as follows,

$$i(\alpha_s + n\lambda_1)A_{sl}^{(n)} + \frac{dA_{s2}^{(n)}}{dx_2} + i(\beta_s + n\lambda_3)A_{s3}^{(n)} = 0, \quad (12a)$$

$$i[-(\omega_s + n\omega_g) + (\alpha_s + n\lambda_1)(1 - x_2^2)]A_{s1}^{(n)} - 2x_2A_{s2}^{(n)} + i(\alpha_s + n\lambda_1)A_{sp}^{(n)} + \frac{1}{\text{Re}} \left[ (\alpha_s + n\lambda_1)^2 + (\beta_s + n\lambda_3)^2 - \frac{d^2}{dx_2^2} \right] A_{s1}^{(n)} + \frac{\varepsilon_w}{2} \Phi_1^{(n)} = 0, \quad (12b)$$

$$i[-(\omega_s + n\omega_g) + (\alpha_s + n\lambda_1)(1 - x_2^2)]A_{s2}^{(n)} + \frac{dA_{sp}^{(n)}}{dx_2} + \frac{1}{\text{Re}} \left[ (\alpha_s + n\lambda_1)^2 + (\beta_s + n\lambda_3)^2 - \frac{d^2}{dx_2^2} \right] A_{s2}^{(n)} + \frac{\varepsilon_w}{2} \Phi_2^{(n)} = 0, \quad (12c)$$

$$i[-(\omega_s + n\omega_g) + (\alpha_s + n\lambda_1)(1 - x_2^2)]A_{s3}^{(n)} + i(\beta_s + n\lambda_3)A_{sp}^{(n)} + \frac{1}{\text{Re}} \left[ (\alpha_s + n\lambda_1)^2 + (\beta_s + n\lambda_3)^2 - \frac{d^2}{dx_2^2} \right] A_{s3}^{(n)} + \frac{\varepsilon_w}{2} \Phi_3^{(n)} = 0. \quad (12d)$$

The quantity  $\Phi_l^{(n)}$ , with  $l = 1, 2, 3$ , that multiplies  $\varepsilon_w$  in (12) is an abbreviation for the following expression,

$$\begin{aligned} \Phi_l^{(n)} = & \left[ i\hat{u}_1[\alpha_s + (n-1)\lambda_1] + i\hat{u}_3[\beta_s(n-1)\lambda_3] + \hat{u}_2 \frac{d}{dx_2} \right] A_{sl}^{(n-1)} \\ & + \left[ i\hat{u}_1^*[\alpha_s + (n+1)\lambda_1] + i\hat{u}_3^*[\beta_s(n+1)\lambda_3] + \hat{u}_2^* \frac{d}{dx_2} \right] A_{sl}^{(n+1)} \\ & + \left[ i\lambda_1 A_{s1}^{(n-1)} + i\lambda_3 A_{s3}^{(n-1)} + A_{s2}^{(n-1)} \frac{d}{dx_2} \right] \hat{u}_l \\ & + \left[ -i\lambda_1 A_{s1}^{(n+1)} - i\lambda_3 A_{s3}^{(n+1)} + A_{s2}^{(n+1)} \frac{d}{dx_2} \right] \hat{u}_l^*. \end{aligned} \quad (13)$$

To bring out the differences between studies of the stability of the flow with a wave excited wall and the classical problem more clearly, it is meaningful to rewrite the ordinary differential equations for the complex amplitude velocity functions  $A_{sl}^{(n)}$  in a form in which the contribution of the Orr-Sommerfeld part and of the additions can be more easily recognised. We therefore subject (12) to the following steps:

Step 1. Multiplication and addition:  $(\alpha_s + n\lambda_1) + (\beta_s + n\lambda_3)$ .  
(12b) (12d)

Step 2.  $(d/dx_2)$  (outcome of step 1)  $- i[(\alpha_s + n\lambda_1)^2 + (\beta_s + n\lambda_3)^2]$ .

Step 3. Multiplication and subtraction:  $(\beta_s + n\lambda_3) - (\alpha_s + n\lambda_1)$ .  
(12b) (12d).

The resulting equation after steps 1 and 2 above is (15) below where the abbreviation  $k_n$  stands for

$$k_n^2 = (\alpha_s + n\lambda_1)^2 + (\beta_s + n\lambda_3)^2. \quad (14)$$

$$\frac{d}{dx_2} \left[ [(\omega_s + n\omega_g) - (\alpha_s + n\lambda_1)(1 - x_2^2)] \frac{dA_{s2}^{(n)}}{dx_2} \right] - (\alpha_s + n\lambda_1) \frac{d}{dx_2} [2x_2 A_{s2}^{(n)}] +$$

$$k_n^2 [-(\omega_s + n\omega_g) + (\alpha_s + n\lambda_1)(1 - x_2^2)] A_{s2}^{(n)} + \frac{1}{\text{Re}} \left[ 2ik_n^2 \frac{d^2}{dx_2^2} - i \frac{d^4}{dx_2^4} - ik_n^4 \right] A_{s2}^{(n)} + \frac{\varepsilon_w}{2} \Psi^{(n)} = 0. \quad (15)$$

The quantity  $\Psi^{(n)}$  which is the multiplier of  $\varepsilon_w$  in (15) is a sum of thirteen terms as follows.

$$\Psi^{(n)} = \sum_{j=1}^{13} \Psi_j^{(n)}. \quad (16)$$

The  $\Psi_j^{(n)}$ ,  $j = 1, 2, 3, \dots, 13$  are listed in Selvarajan & Vasanta Ram (1991).

The outcome of step 3 is (18) below, where  $\Omega^{(n)}$  stands for:

$$\Omega^{(n)} = (\beta_s + n\lambda_3) A_{s1}^{(n)} - (\alpha_s + n\lambda_1) A_{s3}^{(n)}. \quad (17)$$

$$i[-[\omega_s + n\omega_g] + [\alpha_s + n\lambda_1](1 - x_2^2)] \Omega^{(n)} - 2x_2(\beta_s + n\lambda_3) A_{s2}^{(n)} + \frac{1}{\text{Re}} \left[ k_n^2 \Omega^{(n)} - \frac{\delta^2 \Omega^{(n)}}{dx_2^2} \right] + \frac{\varepsilon_w}{2} \chi^{(n)} = 0. \quad (18)$$

The quantity  $\chi^{(n)}$ , the multiplier of  $\varepsilon_w$  in (18), is a sum of twelve terms as follows,

$$\chi^{(n)} = \sum_{j=1}^{12} \chi_j^{(n)}. \quad (19)$$

The  $\chi_j^{(n)}$ ,  $j = 1, 2, 3, \dots, 12$ , are listed in Selvarajan & Vasanta Ram (1991).

Equations (15) and (18), together with the continuity equation (12a), are the set of equations for the complex amplitude functions of the velocity  $A_{s1}^{(n)}$ ,  $A_{s2}^{(n)}$  and  $A_{s3}^{(n)}$ . The boundary condition on all these quantities is zero at  $x_2 = \pm 1$  so that the problem reduces to an eigenvalue problem requiring solution of the dispersion relation (11) for the (complex) frequency  $\omega_s$  (temporal stability problem!).

A cursory inspection of the governing equations shows that for the case of the unexcited wall,  $\varepsilon_w = 0$ , (15) reduces to the Orr-Sommerfeld equation for  $A_{s2}^{(n)}$  and (18) to the Squire equation for  $\Omega^{(n)}$ . For  $\varepsilon_w \neq 0$ , (12a), (15) and (18) form an infinite set of coupled equations. We truncate them at  $n = +1$  and  $-1$ , setting  $A_{sl}^{(n)}$  to zero for all  $n > 1$  and  $n < -1$ . We then have nine unknowns, viz.  $A_{s1}^{(-1)}$ ,  $A_{s2}^{(-1)}$ ,  $A_{s3}^{(-1)}$ ,  $A_{s1}^{(0)}$ ,  $A_{s2}^{(0)}$ ,  $A_{s3}^{(0)}$ ,  $A_{s1}^{(+1)}$ ,  $A_{s2}^{(+1)}$  and  $A_{s3}^{(+1)}$ . The nine equations are obtainable by writing (12a), (15) and (18) for  $n = -1, 0$  and  $+1$ , setting  $A_{sl}^{(n)}$  to zero for all  $n > 1$  and  $n < -1$ .

## 5. Outline of the solution procedure

The solution procedure, both for the basic flow that involves inhomogeneous boundary conditions, and for the problem of further disturbances which is treated as a temporal eigenvalue problem, is given in some detail in Selvarajan & Vasanta Ram (1991). Here we restrict ourselves to observing that the occurrence of the Orr-Sommerfeld operator in both the problems may be used to advantage in the computational procedures. Whereas the basic flow problem was solved by two methods (vide §2) we used only the spectral collocation method together with an iterative scheme to solve the temporal eigenvalue problem.

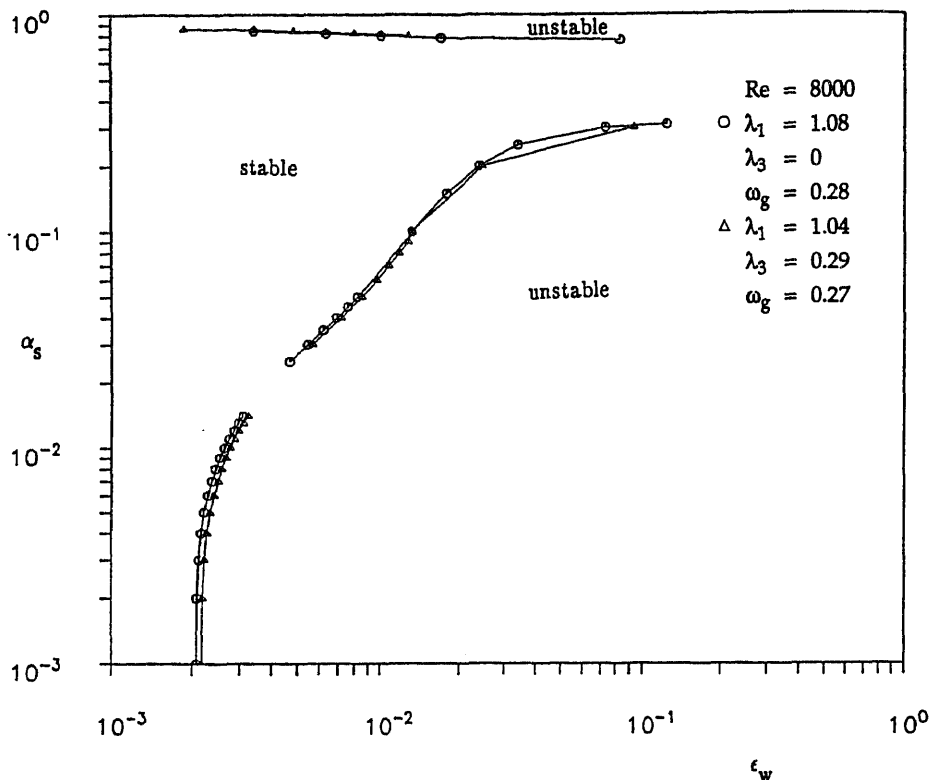


Figure 8. Stability characteristics of the wave-excited channel flow.

## 6. Results

Figure 8 shows the characteristics of the further disturbance in the  $\alpha_s$  and  $\epsilon_w$  plane for a Reynolds number  $Re = 8000$  for two sets of the other parameters whose values are indicated on the figure. Judging from our present results, which should be regarded as preliminary, there are threshold values of the wall excitation amplitude parameter  $\epsilon_w$  which divides the stable from the unstable regions. However, it is interesting to note that there may be disturbances at discrete bands of wave numbers which are damped even at higher amplitudes of excitation. More extensive analytical and computational studies are necessary to substantiate these findings.

A major part of this work was done when one of the authors (SS) spent a period of study at the Ruhr Universität Bochum, Germany. SS wishes to thank the Council of Scientific & Industrial Research for leave, the German Academic Exchange Service for the award of a scholarship and Professor K Gersten, Director of the Institut für Thermo- und Fluidodynamik of the Ruhr University, Bochum for hospitality during this period.

## List of symbols

$A_{sl}^{(n)}$	amplitude function of velocity disturbance, complex quantity, see (10);
$A_{sp}^{(n)}$	amplitude function of pressure disturbance, complex quantity, see (10);
$F$	dispersion relation, defined in (11);
$H$	mean semi-channel height;
$k_n$	defined in (14);
$(n)$	mode number, see (10);
$p$	pressure;
$\tilde{p}$	periodic part of pressure, see (3);
$p$	amplitude function of pressure, complex quantity, see (5);
$p_s$	pressure disturbance, defined in § 3, $l = 1, 2, 3$ , see (8);
$Re$	Reynolds number, $Re = U_0 H / \nu$ ;
$t$	time coordinate;
$u_l$	velocity component in direction $x_l$ , $l = 1, 2, 3$ ;
$\tilde{u}_l$	periodic part of velocity component in $x_l$ -direction, $l = 1, 2, 3$ , see (3);
$\hat{u}_l$	amplitude function of velocity $\tilde{u}_l$ , complex quantity, see (5);
$u_{sl}$	velocity disturbance, defined in § 3, $l = 1, 2, 3$ , see (8);
$U_0$	mean centre line velocity;
$x_l$	spatial coordinates, $l = 1, 2, 3$ ;
$y_w$	wall motion, defined in (2);
$\alpha_s$	wave number of disturbance in $x_1$ -direction, see (10c);
$\beta_s$	wave number of disturbance in $x_3$ -direction, see (10c);
$\varepsilon_s$	amplitude parameter of disturbance, defined in § 3, see (8);
$\varepsilon_w$	amplitude parameter of wall motion, see (2);
$\lambda_1$	wave number in $x_1$ -direction, see (2);
$\lambda_3$	wave number in $x_3$ -direction, see (2);
$\nu$	kinematic viscosity;
$\Phi_l$	function defined in (13), complex quantity, $l = 1, 2, 3$ ;
$\chi^{(n)}$	function defined in (19);
$\Psi^{(n)}$	function defined in (15);
$\omega_g$	frequency, real quantity, see (2);
$\omega_s$	frequency of disturbance, complex quantity, see (10c);
$\Omega^{(n)}$	function defined in (17);
$l, i, j$	indices;
*	as superscript, denotes complex conjugate;
.	above a particular symbol denotes differentiation of that parameter with respect to $x_2$ .

## References

- Bender C M, Orszag S A 1984 *Advanced mathematical methods for scientists and engineers* (New York: McGraw Hill)
- Gottlieb D, Hussaini M Y, Orszag S A 1984 Theory and application of spectral methods. In *Spectral methods for partial differential equations* (eds.) D Gottlieb, M Y Hussaini, S A Orszag (New York: SIAM)
- Liepmann H W, Narasimha R (eds.) 1988 Turbulence management and relaminarization. *IUTAM Symposium, Bangalore, India 1987* (Berlin: Springer Verlag)



- Selvarajan S, Vasanta Ram V 1991 Dynamical characteristics of wave excited channel flows. Report 162 of the Institut für Thermo- und Fluidodynamik, Ruhr Universität, Bochum
- Scott M R, Watts H A 1977 Computational solution of linear two-point boundary value problems via orthonormalization. *SIAM J. Numer. Anal.* 14: 40-70
- Van Dyke M 1975 *Perturbation methods in fluid mechanics* (Stanford, CA: Parabolic Press)

## Detailed study of complex flow fields of aerodynamical configurations by using numerical methods

A DAS

Department of Numerical Aerodynamics, Institute SM-EA, DLR,  
Braunschweig, Germany

**Abstract.** The mathematical physics of fluid flow in a compressible medium, leads to nonlinear partial differential equations or their equivalent integral versions. For the solution of these equations one has generally to resort to numerical methods using mostly finite difference or finite volume schemes, which are well established now. These field methods are very suitable for studying the physical features of complex flows. The present paper gives at first a short sketch of the numerical procedure and thereafter goes into the detailed analysis of the flow fields of delta wings, double-delta wings, delta shaped wing-canard combinations and space vehicles. Further examples include long span wings and wing-bodies at supercritical onflows, flows around propellers and rotors and finally some unsteady flows. The examples cited are selected topics from the extensive studies undertaken in the department of numerical aerodynamics of the DLR in Braunschweig in the course of the last few years.

**Keywords.** Complex flow fields; aerodynamical configurations; numerical aerodynamics.

### 1. Introduction

For the study of complex flows in compressible media it is essential to work with the field solutions of Navier–Stokes equations, or else with the Euler equations when the effects of viscosity and heat conductivity are small, especially at high Reynolds number of the flow. In both cases one has to solve nonlinear partial differential equations, generally by resorting to numerical methods. These methods are well established now and are widely implemented, mostly by using finite-difference or finite-volume schemes. The efficiency of a numerical method is measured by three criteria – good accuracy with robustness, acceptable computing time and easy applicability to complex flow fields. Since the numerical methods for solving nonlinear partial differential equations (Courant *et al* 1928, 1952; Lax & Wendroff 1966) have been established, some basic schemes for solving the Euler equations were introduced thereafter (Beam & Warming 1976; Steger & Warming 1979; Pulliam & Steger 1980; Jameson *et al* 1981; Roe 1981; Whitfield & Janus 1984; Van Leer 1985; Eberle 1987), and numerical schemes for solving Reynolds-averaged Navier–Stokes equations have

also been developed (Beam & Warming 1978; MacCormack 1982, 1985; Jameson 1985; Pan & Lomax 1986; Van Leer *et al* 1987). These basic methods needed however considerable efforts for improved efficiency with respect to the three criteria as stated above and the contributions on this line in recent years have been very significant (some selected papers being Radespiel & Kroll 1985; Kroll & Jain 1987; Rossow 1987, 1989, 1991; Kroll *et al* 1989, 1991; Kroll & Rossow 1990; Radespiel & Kroll 1990; Radespiel *et al* 1990; Blazek *et al* 1991; Swanson & Radespiel 1991; Blazek 1992). For solving the Reynolds-averaged Navier-Stokes equations for turbulent flows it is essential to model the turbulent exchanges producing viscous stresses. A unified and simple approach is the classical eddy-viscosity modelling, which is now being put in an extended version as outlined (Baldwin & Lomax 1978; Degani & Schiff 1983; Johnson & King 1984). One prerequisite for improved efficiency of numerical schemes is the generation of surface and field grids of outstanding quality, as regards smoothness, orthogonality and proper grid fineness especially in regions of high flow gradients. The technique of grid generation is also well established and is applied for producing body fitted grids of various topologies. The basic methods of grid generation and their implementations are discussed (Eriksson 1982; Thompson *et al* 1985; Schwarz 1986; Sonar & Radespiel 1986; Radespiel 1988; Sonar 1989; Findling & Herrmann 1991; Pahlke & Kroll 1991; Rossow & Ronzheimer 1991).

The numerical solutions of Euler and Navier-Stokes equations yield abundant field data in the physical domain around a moving aerodynamical configuration, which can be used for analysing the main features of the flow, both over the boundary surface and in the surrounding space. Such a study is of great importance for complex flow fields involving vortical flows, shock waves and shock boundary layer interactions.

It is well known that slender delta wings moving with moderate to high angles of incidence produce spiralling vortical flows over their upper surface thus causing an appreciable additional lift force, which is termed vortex-lift. This leads to improved aerodynamic properties of the delta-wing aircraft. The flow in the spiralling vortices is complex, being characterized by two special features – loss of total pressure in the spiralling flows and the breakdown of the vortices at high angles of incidence. For some years extensive efforts have been made to study the physics of the vortical flows and to find means for stabilizing them to still higher angles of incidence. Some selected contributions on these lines are cited in the literature (Eriksson & Rizzi 1983; Murman & Rizzi 1986; Rizetta & Shang 1986; Newsome & Kandil 1987; Kumar & Das 1988; Longo 1988, 1992; Raj *et al* 1988; Scherr & Das 1988; Krause & Liu 1989; Agrawal *et al* 1990; Hilgenstock 1990; Hilgenstock & Vollmers 1990; Longo & Das 1990; Das 1991; Das & Longo 1994).

In case of space vehicles using round-nosed wing-body configuration one has to deal with complex flows at subsonic to hypersonic speed regimes involving vortical flows at low speeds and multi-shock flow fields in supersonic and hypersonic velocities. Some interesting studies on these configuration have been made (Risk & Chaussee 1981; Chaussee *et al* 1984; Pfitzner & Weiland 1987; Schöne *et al* 1990, 1991; Radespiel & Swanson 1991; Schöne & Bidault 1991).

Modern transport aircrafts operating with transonic cruising speeds possess long span wings with supercritical aerofoils, thus producing supersonic zones on the upper surface with shock waves and shock induced boundary layer separation. Besides these the wings of a complete aircraft experience interference effects of the fuselage and jet engine nacelles. Again the field solutions using Navier-Stokes and Euler equations are

well suited to study these effects (Chaussee 1986; Deese & Agarwal 1987; Volpe & Jameson 1988; Radespiel 1989; Rossow & Ronzheimer 1991; Rudmik 1991; Longo 1992; Rossow *et al* 1992; Wichmann 1992).

In order to optimize the shapes of propeller blades, propfans and helicopter rotors both in regard to aerodynamics and aeroacoustics it is essential to analyse their flow fields and evaluate the load distributions quite in detail. The outer parts of the blades operating at transonic speeds need to have supercritical aerofoil-shapes, as they are involved in producing supersonic zones and shock waves. Recent investigations (Bober *et al* 1983; Deese & Agarwal 1988; Kroll 1989) report the advancements in numerical aerodynamics.

Further topics of numerical study are the unsteady flow fields as are involved due to oscillating or plunging wings and due to helicopter rotors in forward motion, producing supercritical flow conditions with oscillating shock waves on the upper surface. Some preliminary studies using numerical solutions of Euler equations are undertaken (Whitfield *et al* 1987; Nixon 1989; Carstens 1990; Lin & Pahlke 1991).

In order to validate the results of the numerical solutions of the Euler and Navier-Stokes equations it is necessary to have enough experimental data. With this aim extensive wind-tunnel measurements have been carried out in recent years as are reported (Maynard & Murphy 1950; Caradonna & Phillippe 1976; Bornemann & Surber 1978; Schmitt & Charpin 1979; Caradonna & Tung 1981; Brennenstuhl & Hummel 1982; Lambourne 1982; Redeker *et al* 1987; Drougge 1988; Hummel 1988; Esch 1989; Radespiel & Quast 1989; Bergmann *et al* 1990; Elsenaar & Hoijemakers 1990; Oelkar 1990; Goodard *et al* 1991).

In the present paper some essential features of complex flows around wings, wing-bodies, propellers and rotors as obtained from numerical solutions of Euler- and Navier-Stokes equations are illustrated. The underlying physics helps to understand many of the findings already known from experimental investigations.

## 2. Basic equations of flow fields and their numerical simulation and solution

The disturbance fields arising from the motion of wings and bodies in a compressible viscous medium are adequately described by the Navier-Stokes equations, being based on the conservation laws of mass, momentum and energy in an elementary volume moving with the coordinate system. If the flow is concerned with turbulent exchanges of momentum, a major task is to model these viscous stresses in order to have a complete formulation of the mathematical physics. The usual procedure is to use the classical eddy-viscosity modelling or an equivalent kinematic-dynamical relation.

### 2.1 The field equations in viscous and nonviscous medium

For a fluid medium at standard pressure and temperature having negligible body- or external field-forces the physics of the flow can be fully described by the Navier-Stokes equations in the following form:

$$\frac{D(\rho \mathbf{V})}{Dt} + (\rho \mathbf{V}) \operatorname{div} \mathbf{V} + \operatorname{grad} p - \{\operatorname{div} \boldsymbol{\sigma}_v\} = 0,$$

$$\frac{D(\rho E)}{Dt} + (\rho E) \operatorname{div} \mathbf{V} + \operatorname{div}(p \mathbf{V}) - \{\operatorname{div}(k \operatorname{grad} T) + \operatorname{div}(\mathbf{V} \cdot \boldsymbol{\sigma}_v)\} = 0, \quad (1)$$

where

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{V} \cdot \operatorname{grad}; \quad E = e + (V^2/2),$$

$$\mathbf{V} = iu + jv + kw; \quad \boldsymbol{\sigma}_v \equiv \text{stress tensor.}$$

The set of equations forming (1) contains the unknowns  $u, v, w, p, \rho, T$  and the viscous stress tensor  $\boldsymbol{\sigma}_v$ , thus having more unknowns than the number of equations. An auxiliary relation can be made use of by including the equation of state, which reads

$$p = \rho RT = (\kappa - 1) \rho [E - (u^2 + v^2 + w^2)/2]. \quad (2)$$

While the components of the stress tensor  $\boldsymbol{\sigma}_{vt}$  for laminar flows are easily modelled when the molecular coefficient of viscosity  $\mu_m$  is known, the corresponding stress tensors  $\boldsymbol{\sigma}_{vt}$  for turbulent flows can be based on the Reynolds averaged turbulent stresses expressed as:

$$\sigma_{ii}|_t \equiv \tau_{ii}|_t = -\rho \overline{u_i'^2} \ll p, \quad (\text{normal stress})$$

$$\sigma_{ij}|_t = \tau_{ij}|_t = -\rho \overline{u_i' u_j'}, \quad (\text{tangential stress}) \quad (3)$$

where  $u'_i, u'_j$  etc. are the turbulent fluctuations of the velocities  $u, v, w$ .

Using the concept of eddy-viscosity these two stress-terms can be written in general as

$$\rho \overline{u_i' u_j'} = \mu_t \left[ \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right] - \frac{2}{3} \mu_t \frac{\partial u_k}{\partial x_k} \delta_{ij}, \quad (4)$$

with

$$\mu_t = \rho \bar{l} \bar{v} = \rho \bar{l}^2 |\bar{\omega}|,$$

where  $\bar{l}$  denotes the length scale and  $\bar{v}$  the velocity scale of the turbulent fluctuations, while  $|\bar{\omega}| = (\gamma_i^2 + \gamma_j^2 + \gamma_k^2)^{1/2}$  is the vorticity in the viscous region. Thus (4) has the same expression as for laminar flows with  $\mu_m$  replaced by  $\mu_t$ , which can be determined by using the Baldwin-Lomax (1978) model. The molecular viscosity  $\mu_m$  is yielded by the classical Sutherland formula.

The coefficient of heat conduction  $k$  in the (1) is given by

$$k = \{(\mu_m/\operatorname{Pr}_l) + (\mu_t/\operatorname{Pr}_t)\} \bar{c}_p, \quad (5)$$

where  $\operatorname{Pr}_l$  and  $\operatorname{Pr}_t$  are the Prandtl numbers in laminar and turbulent flows.

For analysing some details of the flow fields containing spiralling flows and vortices it is often convenient to work with the momentum equation in Lambs' version, which reads

$$\frac{D\mathbf{V}}{Dt} = \frac{\partial \mathbf{V}}{\partial t} + \operatorname{grad} \frac{V^2}{2} + \boldsymbol{\gamma} \times \mathbf{V} = -\frac{\operatorname{grad} p}{\rho} + \frac{\operatorname{div} \boldsymbol{\sigma}_v}{\rho} \quad (6)$$

where  $\sigma = -pI + \sigma_v$  denotes the stress dyad, with  $\sigma_v$  as viscous stress tensor. Using the second law of thermodynamics (6) can be rewritten in the Lamb-Crocco version yielding

$$(\partial V / \partial t) + \text{grad}(V^2/2) + \gamma \times V = T \text{grad } s - \text{grad } h + (\text{div } \sigma_v / \rho), \quad (7)$$

$$(\partial V / \partial t) + \gamma \times V = -\text{grad } h_0 + T \text{grad } s + (\text{div } \sigma_v / \rho). \quad (8)$$

Furthermore, it may be necessary to use cylindrical coordinates for evaluating the terms of (6) to (8), thus defining

$$\begin{aligned} \text{grad} &= i \frac{\partial}{\partial r} + j \frac{1}{r} \frac{\partial}{\partial \vartheta} + k \frac{\partial}{\partial x}, \\ \text{div } \sigma_v &= \frac{1}{r} \left\{ \frac{\partial(\sigma_{vr} \cdot r)}{\partial r} + \frac{\partial(\sigma_{v\vartheta} \cdot r)}{\partial \vartheta} + \frac{\partial(\sigma_{vx} \cdot r)}{\partial x} \right\}, \end{aligned} \quad (9)$$

with  $\gamma = i\gamma_r + j\gamma_\vartheta + k\gamma_x$  and  $V = iv_r + jv_\vartheta + kv_x$ .

## 2.2 The Euler equations of a flow field in a perfect medium

For flows in a compressible medium with vanishingly small effects of viscosity and heat conductivity, the terms in the brackets of (1) can be neglected. Using the relation of the first equation in the other two (1) can be expressed in a reduced version, which is commonly known as the classical set of Euler equations for a flow in a perfect medium. They read

$$\begin{aligned} \frac{D\rho}{Dt} + \rho \text{div } V &= 0, \\ \rho \frac{DV}{Dt} + \text{grad } p &= 0, \\ \rho \frac{DE}{Dt} + \text{div}(pV) &= 0, \end{aligned} \quad (10)$$

with  $p = (\kappa - 1)\rho e = \rho RT$ , thus having six equations for the six unknowns  $\rho, u, v, w, p$  and  $T$ .

The field equations described in this section are well suited to study complex flow fields as depicted in figure 1 and are applicable for all flow regimes from subsonic to hypersonic velocities.

## 2.3 The generation of field grids around aerodynamical configurations

The numerical field methods using finite difference or finite volume formulations need suitable field grids around the moving body – the grid spacings should conform to the physical requirements of good flow resolution in regions where high flow gradients are expected and must enable the capture of flow details in the viscous layer close to the body surface. In contrast to this, the grids near the far field boundary can be sparse by having wider stretchings. The quality of the field grids is measured by the smoothness and primarily by the orthogonality to each other and also to the frictional

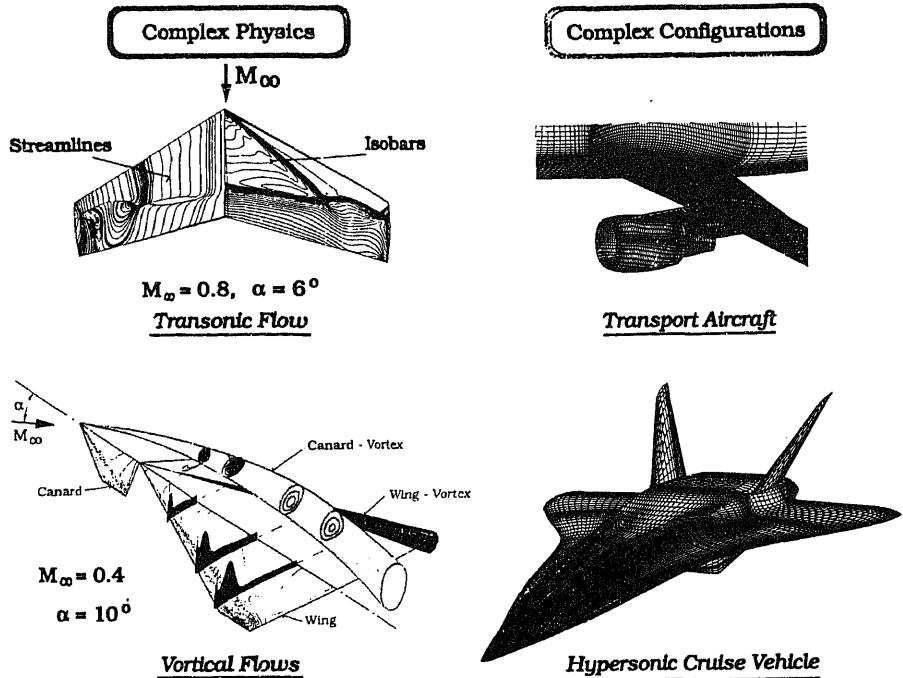


Figure 1. Some complex flow fields in subsonic to hypersonic speed regimes.

surface. The grid topologies to be used depend usually on the nature of the body geometry – they being commonly classified under the notations O-, C-, and H-grids for a given plane, thus leading to the combinations O-O, C-O, C-H etc. for two orthogonal planes, as are shown in figure 2. The grid generation follows three essential steps comprising the following.

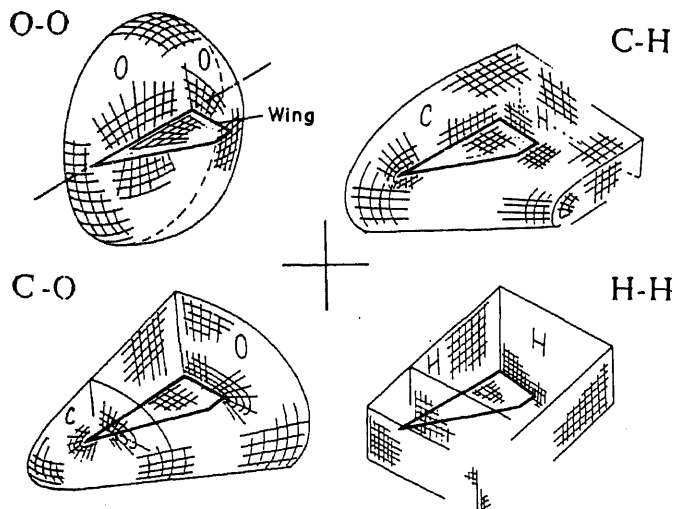


Figure 2. Standard grid topologies for numerical solution of the field equations of bodies moving in an unbounded medium.

- Analytical description of the surface geometry of the moving body, by using interpolation techniques based on cubic splines or Bezier-polynomials. A very elegant procedure is the method described as Coon's surface using linear interpolation with specified pairs of opposite edge points and the corner points of the surface boundary. This scheme is based on Lagrange polynomials.
- Generation of surface grids comprising the distribution of grid points of the solid body, the far field and on the inner cuts introduced in the field. It is essential to have grid clusterings in regions, where high flow gradients are expected. This is realized by using suitable stretching functions.
- Generation of field grids between the solid surface and the far field, for which one of the two basic methods can be followed, both being quite well established and widely implemented till now. These are:
  - a) Initial layout of body-fitted curvilinear coordinates based on algebraic transfinite interpolation schemes with subsequent refinement of the grids for smoothness and orthogonality. For achieving this, the spacings in physical- and computational-domains are interrelated by the solution of elliptic system of equations with source terms, being expressed as Poisson equations  $\nabla^2 X^i = P^i$ , where

$$\nabla^2 = \frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2} + \frac{\partial^2}{\partial \zeta^2}.$$

- b) Use of numerical schemes based on biharmonic equations having the expression

$$\nabla^4 X^i = 0,$$

or

$$\nabla^2 \cdot \nabla^2 X^i = 0. \quad (11)$$

Setting  $\nabla^2 X^i = P^i$ , (11) decomposes into two elliptic systems of equations for the whole field

$$\begin{aligned} \nabla^2 X^i &= P^i, \\ \nabla^2 P^i &= 0. \end{aligned} \quad (12)$$

In both the methods for generating field grids one is to prescribe the first grid spacings normal to the solid surface, which will then yield the  $P^i$ -values at the inner boundary. While in method (b) the  $P^i$ -values in the whole field are yielded by the solution of the Laplace equation  $\nabla^2 P^i = 0$ , the method (a) needs an interpolation of the source terms  $P^i$  from the initial data at the solid surface and at the far field, the latter being usually assumed to have zero source strengths.

Generation of surface grids and field grids around a wing-body combination has been depicted in figure 2. Depending on the storage capacity of the computer it is normally necessary to subdivide the grid space in multiblocks and handle the blocks in turn.

### 2.3 Numerical simulation of the field equations and the method of solution

The Navier-Stokes equations describing the physics of flow in a compressible viscous medium, as expressed in (1), can be written in divergence form in the following way:

$$\frac{\partial \rho}{\partial t} + \text{div}(\rho \mathbf{V}) = 0,$$



$$\frac{\partial(\rho V)}{\partial t} + \operatorname{div}(\mathbf{V}; \rho \mathbf{V}) + \operatorname{div}(pI) - \{\operatorname{div} \sigma_v\} = 0,$$

$$\frac{\partial(\rho E)}{\partial t} + \operatorname{div}(\rho E \mathbf{V}) + \operatorname{div}(pV) - \{\operatorname{div}(k \operatorname{grad} T) + \operatorname{div}(\mathbf{V} \cdot \sigma_v)\} = 0, \quad (13)$$

with  $I = (i:i + j:j + k:k)$  as a unit dyad. The divergence of the momentum flux in the second expression contains a dyadic product comprising all the nine components of the flux in the three coordinate directions. For a perfect medium the terms in the curly brackets drop out and (13) reduces then to the Euler equations in divergence form.

For numerical simulation of (13) it is essential to rewrite it in a universal vector form as:

$$(\partial U / \partial t) + \nabla \cdot \chi = 0, \quad (14)$$

where the solution vector  $\mathbf{U}$  and the flux tensor  $\chi$  are defined as

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ \rho E \end{bmatrix}, \quad \chi = \begin{bmatrix} \rho \mathbf{V} \\ \rho u \mathbf{V} + p \mathbf{i}_x + \sigma_{xx} + \tau_{xy} + \tau_{xz} \\ \rho v \mathbf{V} + p \mathbf{i}_y + \sigma_{yy} + \tau_{yx} + \tau_{yz} \\ \rho w \mathbf{V} + p \mathbf{i}_z + \sigma_{zz} + \tau_{zx} + \tau_{zy} \\ \rho E \mathbf{V} + pI + \sigma_v \cdot \mathbf{V} - k \operatorname{grad} T \end{bmatrix}, \quad (15)$$

with the viscous flux tensor

$$\sigma_v \cdot \mathbf{V} = \begin{bmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_{yy} & \sigma_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{bmatrix} \cdot \begin{bmatrix} u \\ v \\ w \end{bmatrix}.$$

In the Cartesian system of coordinates the flux tensor of (15) can be split as:

$$\chi = F \mathbf{i}_x + G \mathbf{i}_y + H \mathbf{i}_z, \quad (16)$$

so that (14) can be rewritten as

$$\frac{\partial U}{\partial t} + \frac{\partial E}{\partial x} + \frac{\partial G}{\partial y} + \frac{\partial H}{\partial z} = 0. \quad (17)$$

For a finite volume element  $\Omega$  enclosed by a surface  $S$ , (14) can be expressed in integral form. If the volume integral of the flux tensor is converted to surface integral by using the Gauss-theorem, one obtains:

$$\frac{\partial \bar{U}}{\partial t} + \frac{1}{\Omega} \int_S \chi \cdot \mathbf{n} dS = 0, \quad (18)$$

where  $\bar{U} = \frac{1}{\Omega} \int_{\Omega} U d\Omega$ , and  $dS$  is a surface element of  $S$  with  $\mathbf{n}$  as its unit outer-normal vector.

In case of small effects of viscosity and heat conductivity the corresponding terms

in (15) can be dropped, thus leading to the integral version of the Euler equations. The Navier–Stokes equations for unsteady flows as expressed in (13) being basically of hyperbolic type need the specification of the following conditions for the solution of a given problem.

- (a) The initial condition as prescribed by the onflow at the time  $t = t_0$ .
- (b) Boundary condition on the solid surface

$$\begin{array}{ll} \text{pressure:} & \partial p / \partial n = 0, \quad \text{in the viscous layer close to the solid boundary;} \\ \text{velocity:} & \mathbf{V} = 0, \quad \text{no slip at the solid surface;} \\ \text{temperature:} & \begin{cases} k(\partial T / \partial n) = 0, & \text{adiabatic condition at the wall;} \\ T = T_w, & \text{isothermal condition at the wall.} \end{cases} \end{array}$$

(In case of inviscid flows the boundary condition at the solid surface simplifies to  $\mathbf{V} \cdot \mathbf{n} = \mathbf{V}_\infty \cdot \mathbf{n} + \Delta \mathbf{V} \cdot \mathbf{n} = \mathbf{V}_s \cdot \mathbf{n}$ , where  $\mathbf{V}_s$  denotes unsteady motion of the body surface).

- (c) Boundary condition at the far field is based on the characteristic relations, so that the propagation of information from inside and outside are properly matched, thus preventing spurious reflections into the enclosed domain.
- (d) Condition of periodicity at the inner cuts used in the solution and also of matching of the flow variables at the block boundaries of multiblock grid topologies.

The numerical simulation of the unsteady Navier–Stokes or Euler equations in conservation law form can be undertaken by converting into a finite difference equation, preferably by using body-fitted coordinates.

If curvilinear coordinates are used based on the coordinate transformations

$$\xi = \xi(x, y, z); \quad \eta = \eta(x, y, z); \quad \zeta = \zeta(x, y, z),$$

then the unit vectors and the flux tensors are to be redefined in the  $\xi, \eta, \zeta$  coordinates, by using the metric coefficients and the determinant  $J$  of the Jacobian matrix.

$$\begin{aligned} \nabla &= \mathbf{i}_\xi \frac{\partial}{\partial \xi} + \mathbf{i}_\eta \frac{\partial}{\partial \eta} + \mathbf{i}_\zeta \frac{\partial}{\partial \zeta}, \\ \tilde{F} &= \left[ F \frac{\partial \xi}{\partial x} + G \frac{\partial \xi}{\partial y} + H \frac{\partial \xi}{\partial z} \right] \frac{1}{J(\xi, \eta, \zeta)}, \\ \tilde{G} &= \left[ F \frac{\partial \eta}{\partial x} + G \frac{\partial \eta}{\partial y} + H \frac{\partial \eta}{\partial z} \right] \frac{1}{J(\xi, \eta, \zeta)}, \\ \tilde{H} &= \left[ F \frac{\partial \zeta}{\partial x} + G \frac{\partial \zeta}{\partial y} + H \frac{\partial \zeta}{\partial z} \right] \frac{1}{J(\xi, \eta, \zeta)}, \end{aligned} \quad (19)$$

and  $\tilde{U} = U/J(\xi, \eta, \zeta)$ , while the volume  $\tilde{\Omega} = \Omega \cdot J(\xi, \eta, \zeta)$ . Hence (17) takes the following form

$$\frac{\partial \tilde{U}}{\partial t} + \frac{\partial \tilde{F}}{\partial \xi} + \frac{\partial \tilde{G}}{\partial \eta} + \frac{\partial \tilde{H}}{\partial \zeta} = 0, \quad (20)$$

thus retaining the conservation law form.

The numerical simulation of (20) in finite-difference form can be undertaken by using explicit or implicit formulations.

$$U_{ijk}^{(n+1)} = U_{ijk}^{(n)} - \Delta t \left\{ \frac{D_-}{\Delta \xi} F_{ijk}^{(n)} + \frac{D_-}{\Delta \eta} G_{ijk}^{(n)} + \frac{D_-}{\Delta \zeta} H_{ijk}^{(n)} \right\}. \quad (21)$$

Implicit scheme

$$\begin{aligned} \tilde{U}_{ijk}^{(n+1)} = U_{ijk}^{(n)} - \frac{\Delta t}{2} \left[ \left\{ \frac{D_0}{\Delta \xi} F_{ijk}^{(n)} + \frac{D_0}{\Delta \eta} G_{ijk}^{(n)} + \frac{D_0}{\Delta \zeta} H_{ijk}^{(n)} \right\} \right. \\ \left. + \left\{ \frac{D_0}{\Delta \xi} F_{ijk}^{(n+1)} + \frac{D_0}{\Delta \eta} G_{ijk}^{(n+1)} + \frac{D_0}{\Delta \zeta} H_{ijk}^{(n+1)} \right\} \right], \quad (22) \end{aligned}$$

with  $(n+1)$  denoting the time step  $t + \Delta t \equiv (n+1)\Delta t$ . While  $D_-$  denotes upwind differencing,  $D_0$  stands for central differences, as are elucidated in figure 3.

For numerical simulation of (18) in finite-volume formulation one can use arbitrary meshes, thus yielding

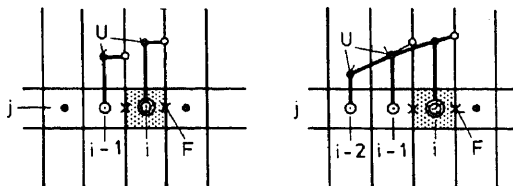
$$\begin{aligned} U_{ijk}^{(n+1)} = U_{ijk}^{(n)} - \frac{\Delta t}{\Omega_{ijk}} \left\{ (\chi \cdot \mathbf{S})_{i+\frac{1}{2},j,k} + (\chi \cdot \mathbf{S})_{i-\frac{1}{2},j,k} + (\chi \cdot \mathbf{S})_{i,j+\frac{1}{2},k} \right. \\ \left. + (\chi \cdot \mathbf{S})_{i,j-\frac{1}{2},k} + (\chi \cdot \mathbf{S})_{i,j,k+\frac{1}{2}} + (\chi \cdot \mathbf{S})_{i,j,k-\frac{1}{2}} \right\}^{(n)}. \quad (23) \end{aligned}$$

Using the flow variables and flux vectors in transformed coordinates  $(\xi, \eta, \zeta)$  as defined in (19) the numerical simulation of the flow equations reduces to the following expression:

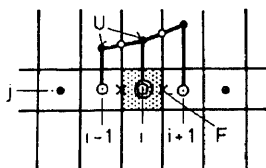
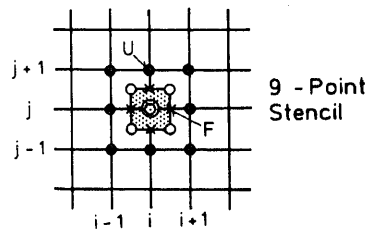
$$\tilde{U}_{ijk}^{(n+1)} = \tilde{U}_{ijk}^{(n)} - \Delta t \left[ \frac{D_0}{\Delta \xi} \tilde{F}_{ijk} + \frac{D_0}{\Delta \eta} \tilde{G}_{ijk} + \frac{D_0}{\Delta \zeta} \tilde{H}_{ijk} \right]^{(n)}. \quad (24)$$

Cell-Centred Schemes

Cell-Vertex Schemes



Upwind Schemes



Central-Average Scheme

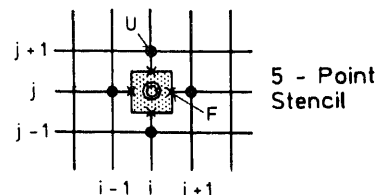


Figure 3. Standard numerical schemes for simulating the equations of flow fields.

Thus (24) reproduces (21) exactly with  $D$  replaced by  $D_0$ , confirming the equivalence of finite-difference and finite-volume methods.

The basic solution scheme of the viscous and nonviscous flow equations, which has been used in the present study, is a finite volume discretization with a Runge-Kutta integration in time, as is described by Jameson *et al* (1981). The original cell-centred code (Radespiel & Kroll 1985) has been extended now to a cell-vertex code (Rossow 1989). Since the finite difference discretization for numerical simulation is based on the central difference scheme, which is insensitive to sensor sawtoothed spatial waviness of the field quantities, that may creep in due to odd-even point decoupling, an artificial dissipative term is added to the equation to damp out the high frequency oscillations. Hence, the numerical approximation to the integral equation (18) as expressed in (13) becomes extended to the form

$$U_{ijk}^{(n+1)} = U_{ijk}^{(n)} - \frac{\Delta t}{\Omega_{ijk}} \{Q_c + Q_v + D\}_{ijk}^{(n)}, \quad (25)$$

where

$$Q_c = \sum_h \delta_h (\chi_c \cdot S), \quad Q_v = \sum_h \delta_h (\chi_v \cdot S), \quad \text{with } (h \equiv i, j, k),$$

denoting the balance of convective flux and of viscous flux, while  $D$  is an artificial dissipative term. For the method of evaluation of the three flux terms of an elemental volume at a grid point  $i, j, k$  of the flow field, one may refer to the details of the discretization procedure discussed by Radespiel & Rossow (1990). In case of Euler equations for inviscid flows the flux term  $Q_v$  in (25) drops out.

For the solution of (25) a five-stage Runge-Kutta time stepping scheme is used in the following way:

$$\begin{aligned} U_{ijk}^{(0)} &\equiv U_{ijk}^{(n)} \\ U_{ijk}^{(r)} &= U_{ijk}^{(0)} - \alpha_r \frac{\Delta t}{\Omega_{ijk}} [Q_c + Q_v + D]_{ijk}^{(r-1)}, \\ U_{ijk}^{(n+1)} &= U_{ijk}^{(5)}, \end{aligned} \quad (26)$$

where  $r = 1$  to 5 and the stage coefficients  $\alpha_r$  are

$$\alpha_1 = 1/4; \alpha_2 = 1/6; \alpha_3 = 3/8; \alpha_4 = 1/2; \alpha_5 = 1.$$

The artificial dissipation  $D^{(r-1)}$  may be evaluated only for the 1st, 3rd and 5th stages of the scheme. The artificial dissipation model is based on the fluctuation of pressure as a sensor (Jameson 1981) defined by

$$\bar{v}_{ijk} = \frac{p_{i-1,j,k} - 2p_{ijk} + p_{i+1,j,k}}{p_{i-1,j,k} + 2p_{ijk} + p_{i+1,j,k}}. \quad (27)$$

For the evaluation of the dissipative operators in regions of normal grid structures and for highly stretched grids in the viscous layer one may refer to the details discussed by Radespiel (1989). The inclusion of the dissipative fluxes leads to the convergence of the procedure to a smooth solution. The stability of the numerical scheme was assured by choosing a CFL-no. of 9.5 both in the inner and outer block.

As steady state solutions are sought for, the use of several accelerating techniques

is allowed to advance the solution. They are successive grid refinements, local time stepping, enthalpy damping, implicit residual averaging and finally the multigrid technique.

### 3. Numerical study of complex flow fields around aerodynamical configurations

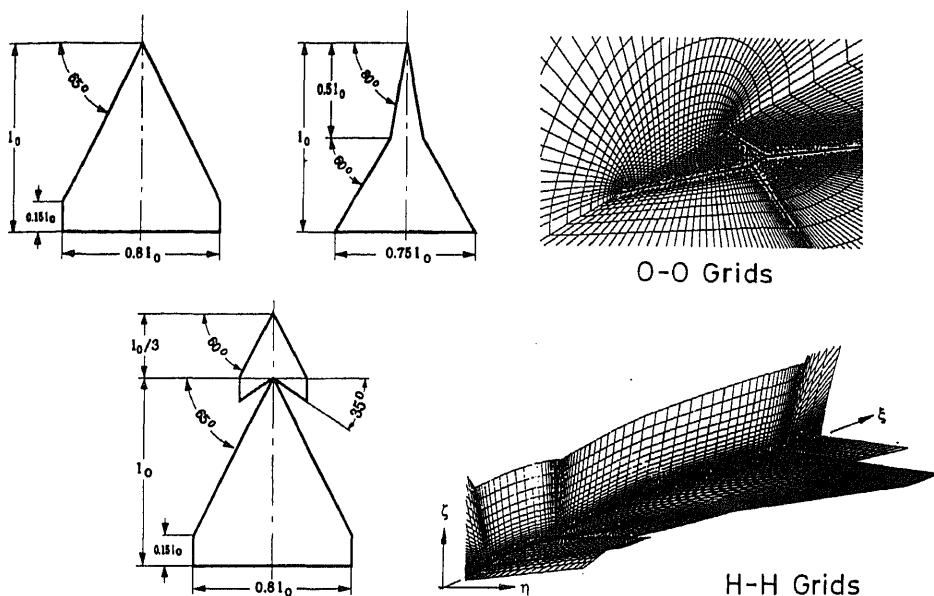
The nonlinear partial differential equations describing the physics of a flow field are inherently complex and the physics itself becomes more complex when vortical flows, shock-waves or shock-viscous layer interactions come into being – especially in transonic and hypersonic speed regimes. Further cases of complex flows arise when the aerodynamical configurations themselves have involved geometries, as in the case of a complete aircraft configuration with wing, fuselage, nacelle and pylon and possibly with contrarotating propfans with highly swept blades. Numerical studies and analysis of such flow fields are becoming more and more common now. While the numerical solution of Euler equations is widely used to study the global features of complex flow fields, more effort is needed to analyse the details of the viscous effects, especially in regions of high flow gradients, which the solution of the Navier–Stokes equations depends on. The following examples will elucidate the findings from a number of interesting studies.

#### 3.1 *Vortical flow fields around delta wings and a delta shaped wing-canard combination*

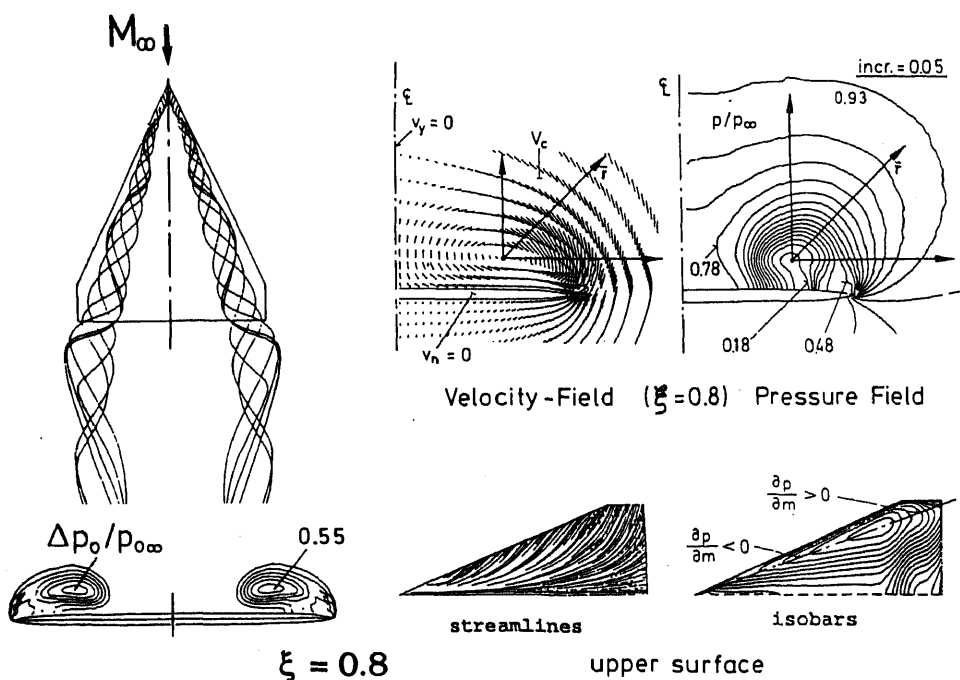
The flow field of a slender delta wing moving in a compressible medium offers an ideal example for the study of complex flows, extended over a speed range from subsonic through transonic to supersonic onflow velocities. This is due to the fact that at moderate to high incidence angles of the wing a large part of the flow field is involved with the formation of vortices spiralling over the upper surface of the wing, thus contributing substantially to an additional vortex-lift which is useful for an outstanding flight performance. The spiralling flow is established by calling in a loss of total pressure thus giving rise to vorticities in the flow. Besides which shock waves and shock viscous-layer interactions are common at high subsonic to supersonic onflow velocities. A further interesting feature is the breakdown of the spiralling vortices at high incidence angles, causing a loss in the vortex-lift. Thus a detailed study of the physics of such flows can lead to many interesting findings.

The delta wings selected for the numerical studies are depicted in figure 4, their having undergone extensive experimental wind-tunnel tests (Brennenstuhl & Hummel 1982; Drouge 1988; Bergmann *et al* 1990; Elsenaar & Hoijemakers 1990; Oelker 1990). While for the delta wings O–O grid topologies have been used, it proved to be simple to provide H–H grid topology for the wing-canard combination and maintain the same grid structure for canard-on and -off configurations.

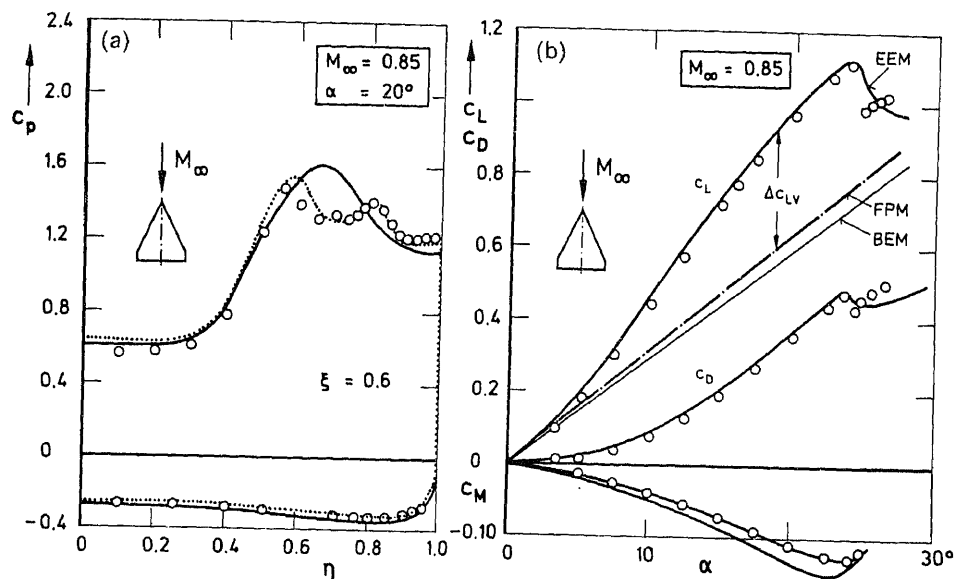
The numerical study with the simple delta wing has been undertaken by using the Euler equations, the solution of which yields all the essential field data which can be plotted and depicted as cross-flow fields or surface distributions of velocities and pressure, as have been shown in figure 5. The trace of the spiralling vortices and the isolines of total pressure at a cross-cut reveal further details. In order to validate the computational data more closely with those of experiments the  $c_p$ -distributions



**Figure 4.** Delta wing configurations for numerical study of vortical flow fields by using Euler- and N-S equations.



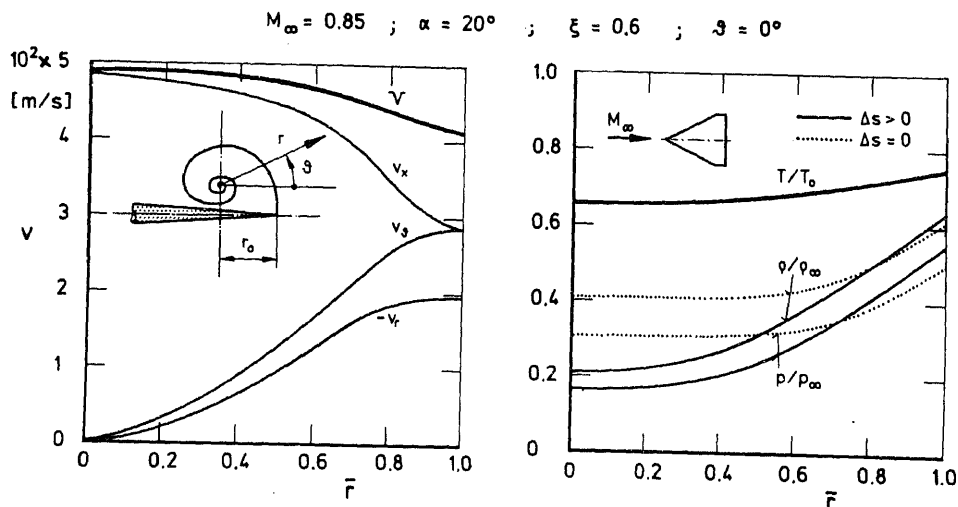
**Figure 5.** Numerical study depicting the cross-flow and surface-flow of a delta wing as yielded by the solution of Euler equations.  $M_\infty = 0.85$ ;  $\alpha = 20^\circ$ .



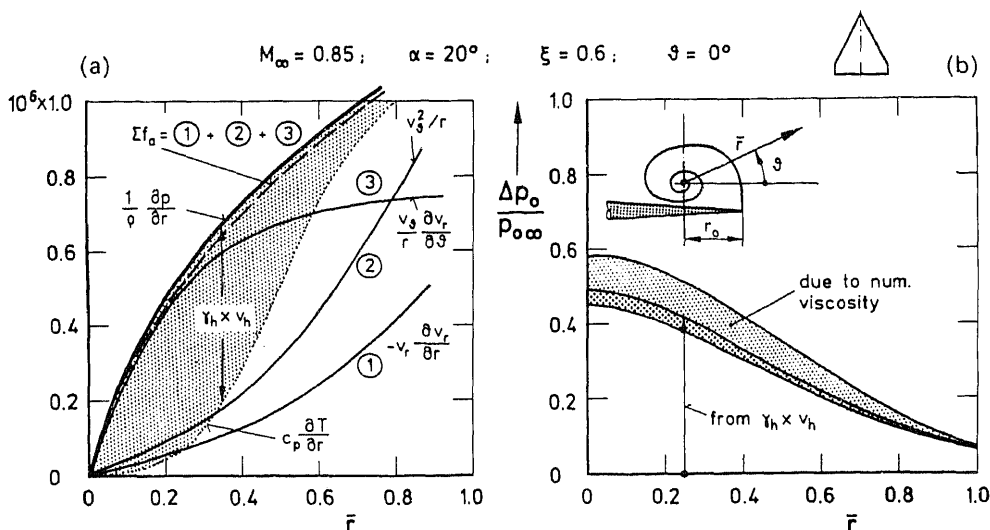
**Figure 6.** Comparison of numerical (Longo & Das 1990) and experimental (Elsenaar & Hoijemakers 1990) values of surface pressures (a) and aerodynamic forces (b) of the delta wing. In (a)  $\circ$  represent experimental values, — are values calculated using the Euler equation and .... are values from the Navier-Stokes equations. In (b)  $\circ$  are experimental values while — are numerical values.

on the wing surface are compared in figure 6 and it was observed that the total lift and drag forces confirm very good agreement.

Having now the field data one can take up interesting analysis on the physics of the flow, regarding the setting in of the spiralling motion with loss of total pressure in it and on the changes in the structure of the vortical flow till its breakdown at



**Figure 7.** Pressure-, temperature- and velocity-distribution in the spiralling vortical-flow of a delta wing.



**Figure 8.** The pressure forces (a) and loss in total pressure (b) required for setting in the spiralling flow.

high incidence (Das 1991). With the imposed boundary conditions on the wing surface, and at the symmetry section of the wing the velocity and pressure distributions in the vortex take their typical runs, as are shown in figure 7 for a cross-flow plane  $\xi = 0.6$ . It is evident that an isentropic pressure distribution corresponding to the velocity  $V$  can by no means bring up the required accelerating forces to establish the spiralling – the flow has to be rotational, thus causing loss in total pressure for adjusting the radial forces which are needed. The nature of the accelerating forces and loss of total pressure along a radial line from the vortex core are depicted in figure 8.

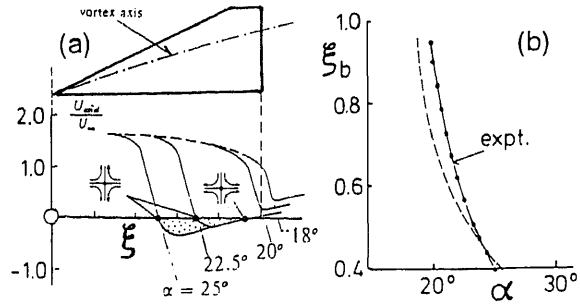
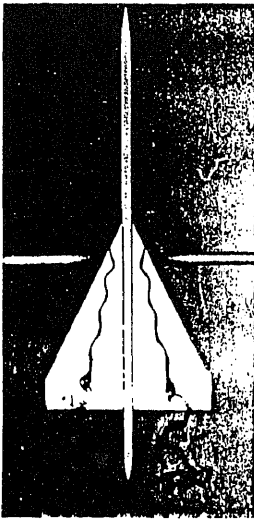
From the upper surface isobars of the delta wing as shown in figure 5 one can observe that an adverse pressure gradient appears in the rear part of the wing, which increases significantly at higher angles of incidence. The loss of total pressure and the adverse pressure gradient in the spiralling flow both increasing with the angle of incidence of the wing leads to the formation of two saddle points along the vortex axis – the one causes a reverse flow and the other contributes to high radial flow outward as has been depicted in figure 9 and is well elucidated (Das 1991). As a result the spiralling structure of the vortex breaks down, now causing a drop in the vortex-lift.

In order to improve the aerodynamic properties of delta wings various planforms have been investigated in the past, – one promising configuration being the strake – or double-delta wing. Numerical study on the strake wing shown in figure 4 was undertaken by using the Navier–Stokes equations and turbulent eddy viscosity modelling. A comparison of the pressure distributions and of the total forces and moments with the experimental values, as are shown in figure 10, confirm the validation of both the results. Further comparisons of the surface flow and cross-flow of the wing as yielded by the numerical and experimental results in figure 11 prove the reliability of the method of calculation. For more examples of such studies one can refer to the cited literature (Das & Longo 1994a).



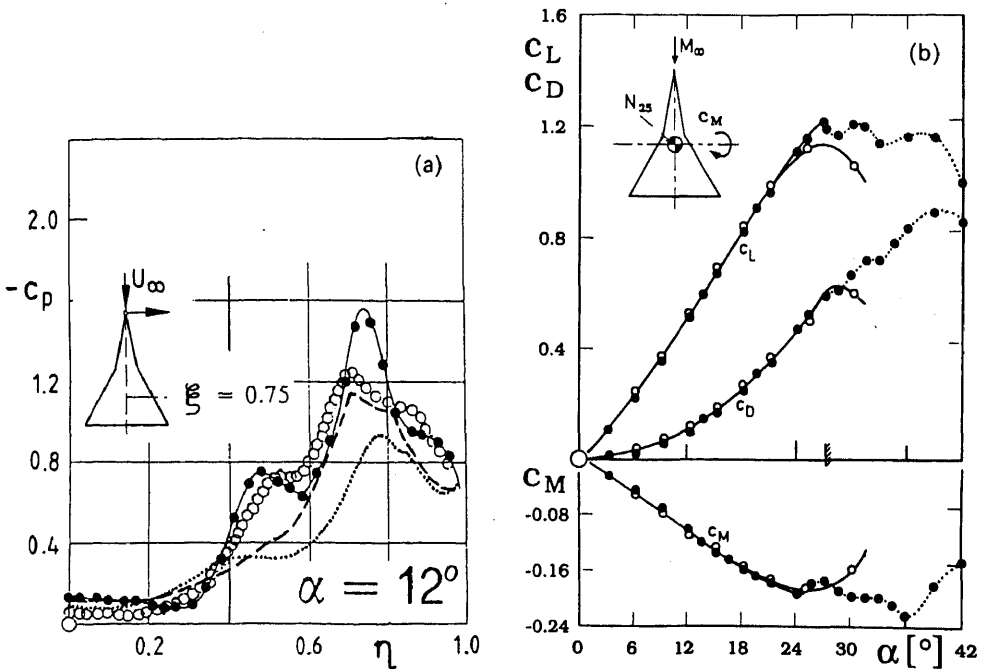
$$\alpha = 19.5^\circ$$

$$\text{Num. Comp.: } M_\infty = 0.4$$

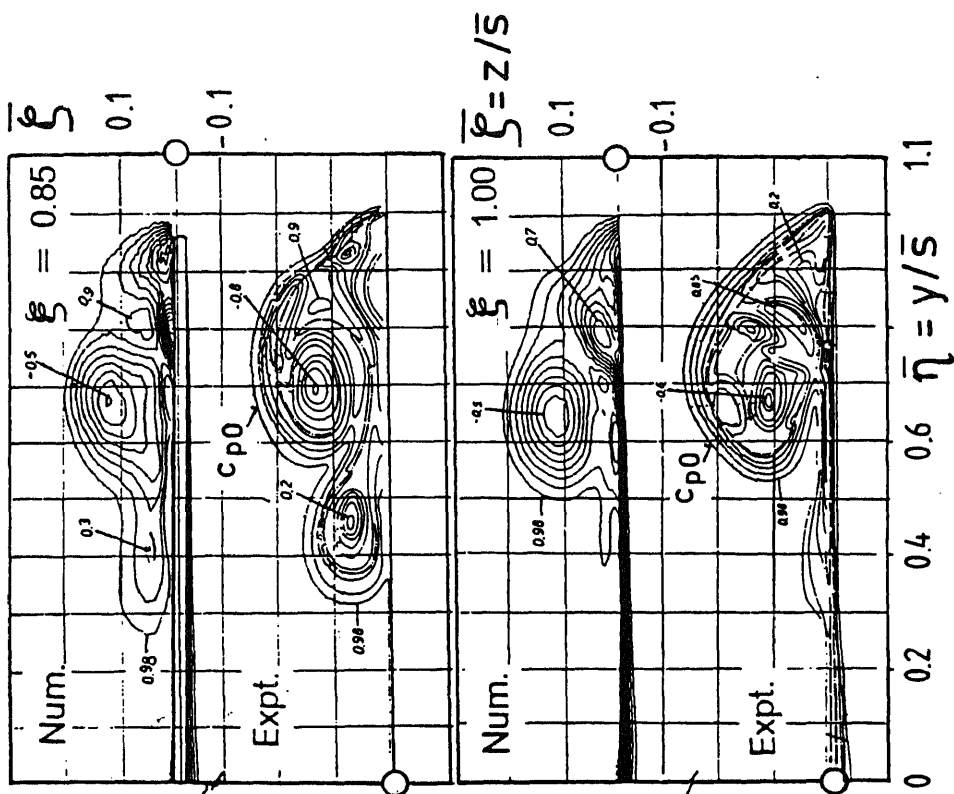


Expt. ( $M_\infty = 0.1$ )

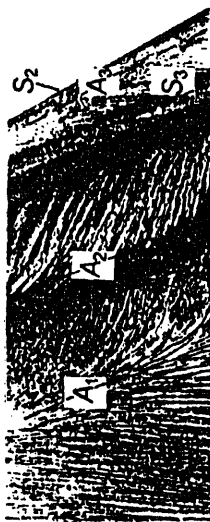
**Figure 9.** Analysis of the flow condition leading to vortex breakdown and loss of vortex-lift. (a) Flow reversal with saddle points and (b) vortex breakdown.



**Figure 10.** Surface pressure and aerodynamic force coefficients of the double-delta wing. (a) Cross-wise  $c_p$  distribution [ $\bullet$  experimental (Brennenstuhl & Hummel 1982);  $\circ$  present – numerical (Das & Longo 1994a); --- numerical (Krause & Liu 1989);  $\cdots$  numerical (Hartwich *et al* 1988)]; (b) forces and moment [ $Re = 1.33 \times 10^6$ ;  $\circ$  numerical ( $M_\infty = 0.3$ );  $\bullet$  experimental ( $M_\infty = 0.1$ , Brennenstuhl & Hummel 1982)]



Expt. ( $M_\infty = 0.1$ ) (Brennenstuhl & Hummel 1982)



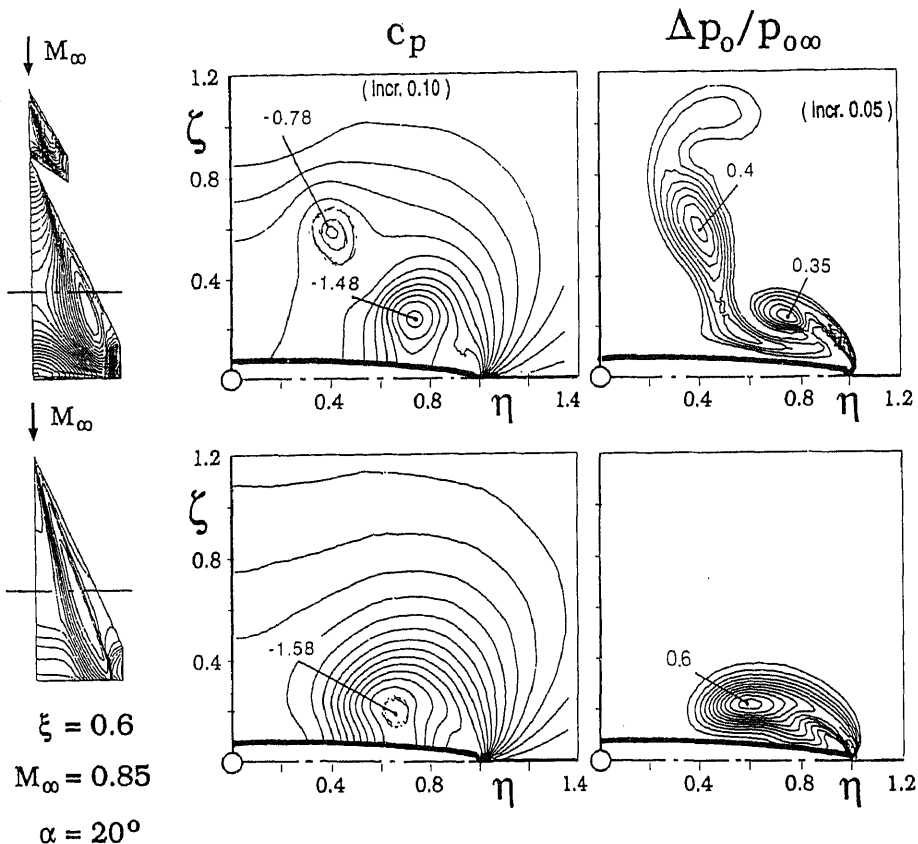
Num. ( $M_\infty = 0.3$ )



A : attachment line  
S : separation line

1 : primary vortex (strake)  
2 : primary vortex (wing)  
3 : secondary vortex (wing)

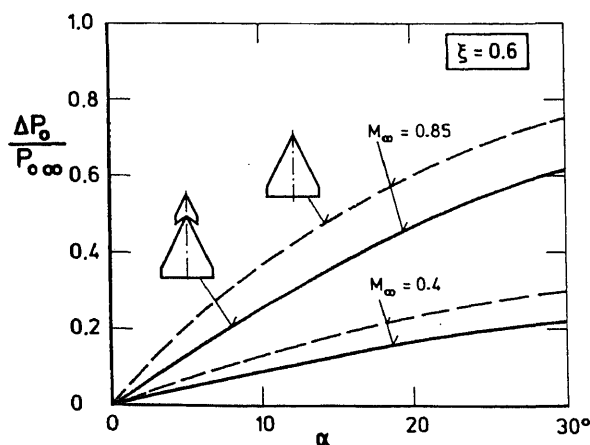
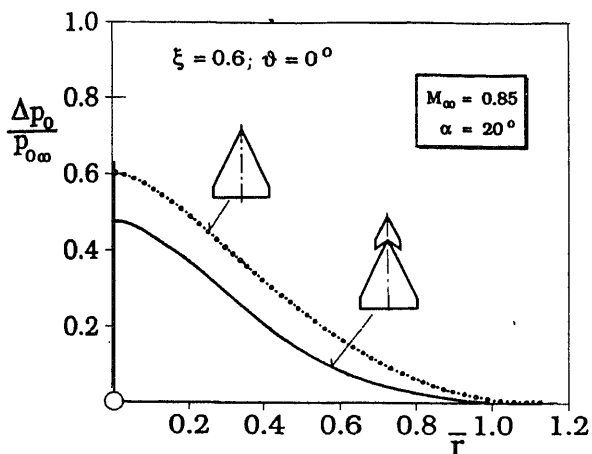
Figure 11. Numerical study of the flow around a double-delta wing by using N-S equations.  $\alpha = 12^\circ$  and  $Re = 1.33 \times 10^6$ .



**Figure 12.** The field quantities in the cross-flow plane of a wing-canard configuration as yielded by the solution of Euler equations.

It is known from wind tunnel measurements as well as flight tests that a delta shaped wing-canard configuration distinguishes itself through outstanding aerodynamic properties, primarily by maintaining its vortex-lift up to high incidence angles. Thus it offers an ideal example to analyse the complex flows due to multi-vortices spiralling over the wing.

So one can take up similar studies as with the simple delta wing already described above. The numerical field data have been plotted and depicted in a similar way, one example being shown in figure 12. The canard imparting downward momentum to the air particles ahead of the wing produces canard-lift and a decrease in the wing-lift due to the downwash created by it. Because of this and due to the canard vortex spiralling over the wing, the wing vortex becomes weaker having less loss in total pressure and also less adverse pressure gradient in the rear region. The loss of total pressure in the vortex core with the canard off and on has been compared in figure 13. Consequently, the spiralling vortex structure and the vortex-lift are maintained without breakdown up to high incidence angles, as is evident from the curves of total forces in figure 14. With a closely coupled canard the drop in the wing-lift is just compensated by the canard-lift; however, the lift curve  $c_L(\alpha)$  continues its rise with the same slope up to higher incidence angles than the wing alone.

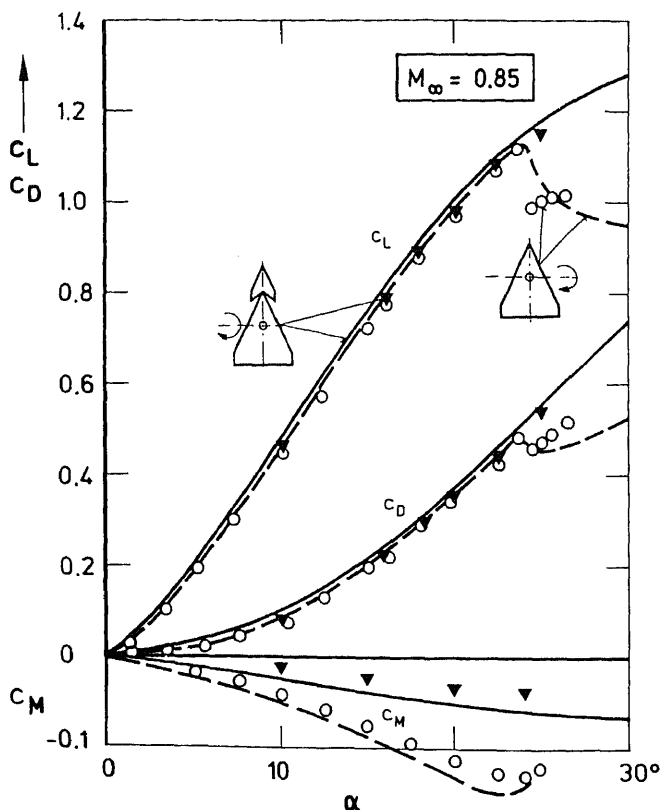


**Figure 13.** Loss of total pressure in the vortex core of a wing-canard configuration with the canard on and off.

Extensive analysis of the physics of the complex flow around this wing-canard combination has been presented in Das & Longo (1994b).

### 3.2 Flow fields around space vehicles

Further studies on flow fields of slender delta wings at subsonic up to supersonic and hypersonic speeds concern the configurations used for space vehicles – the two present examples being the American space-shuttle and the European space-project HERMES, as are shown in figure 15, both have extensive experimental data, as cited in the literature (Bornemann & Surber 1978; Esch 1989; Radespiel & Quast 1989). The recent measurements on a space shuttle model (Radespiel & Quast 1989) comprise detailed investigations of surface flows and pressure distributions, as well as of total forces and moments. In order to reproduce all these the numerical studies were based on the Navier–Stokes equations with eddy-viscosity modelling of the turbulent viscous stresses (Baldwin & Lomax 1978). A comparison of the numerical and experimental results is shown in figure 16 confirming excellent agreement. The details of the numerical method have been discussed by Das & Longo (1994a).

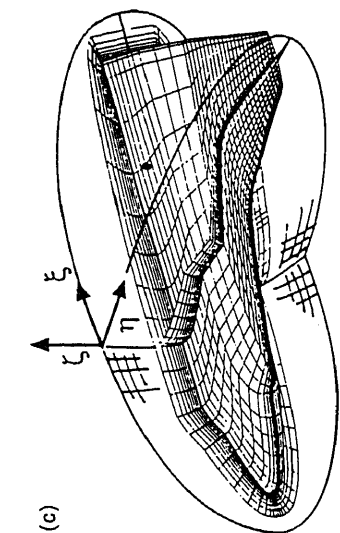


**Figure 14.** Aerodynamic coefficients of total lift, drag and moment of a wing-canard configuration with the canard on and off. [— numerical (Das & Longo 1994a); --- numerical (Longo & Das 1990); ▼ ○ experimental (Elsenaar & Hoijemakers 1990).]

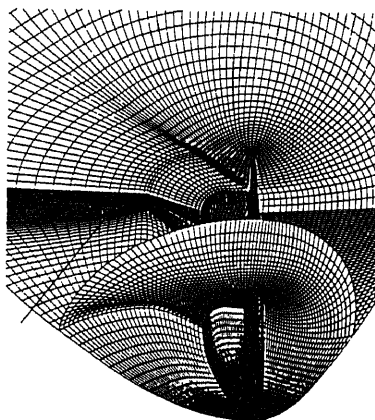
Numerical studies on the HERMES-configuration were undertaken for flight conditions at supersonic Mach number  $M_\infty = 2.5$  to  $8.0$  for symmetrical and unsymmetrical onflows. Because of the complex flows over the wing and body, especially at unsymmetrical onflows, it was essential to base the investigations on the solution of the Euler equations. The field data have been plotted as cross-flow velocities, cross-flow isobars and iso-Mach-lines as well as surface plots of isobars and streamlines. Finally the force coefficients  $c_L$ ,  $c_D$  and  $c_M$  have been evaluated for symmetrical onflow as well as the stability derivatives at unsymmetrical flow conditions, as are depicted in figures 17 and 18. Comparison with the experimental results proves the outstanding reliability of the numerical methods used. For further details on this work one may refer to the paper by Schöne and coworkers (Schöne & Bidault 1991; Schöne *et al* 1991).

### 3.3 Flow fields of large span wings including wing-fuselage and wing-nacelle interference

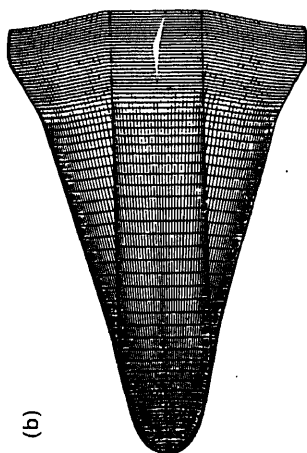
Most of the present transport airplanes with large span wings fly at transonic Mach numbers, such that the onflow velocity to the wings lies in the supercritical regime,



(a)



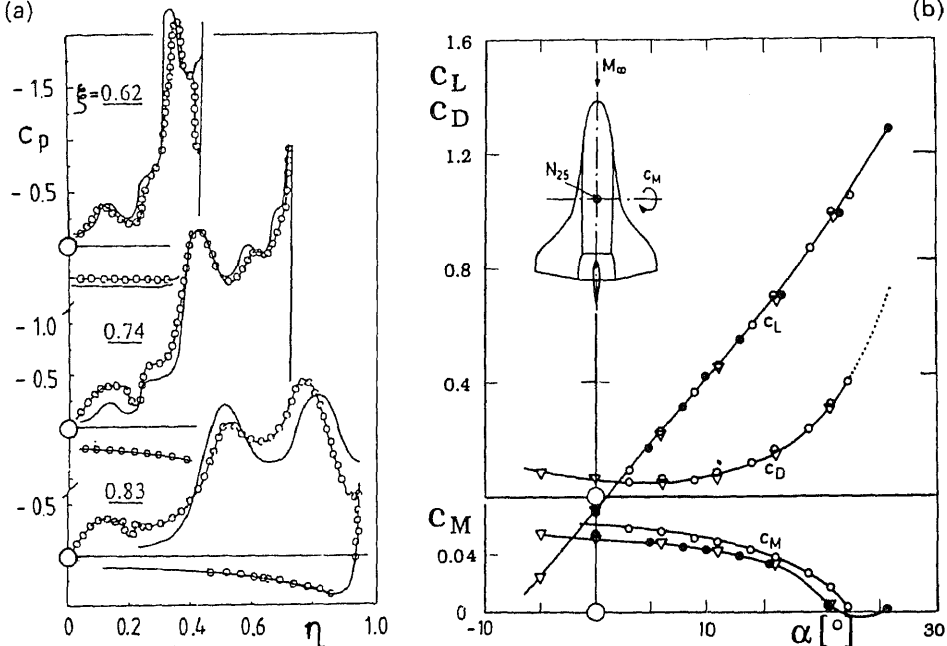
(b)



(c)

(d)

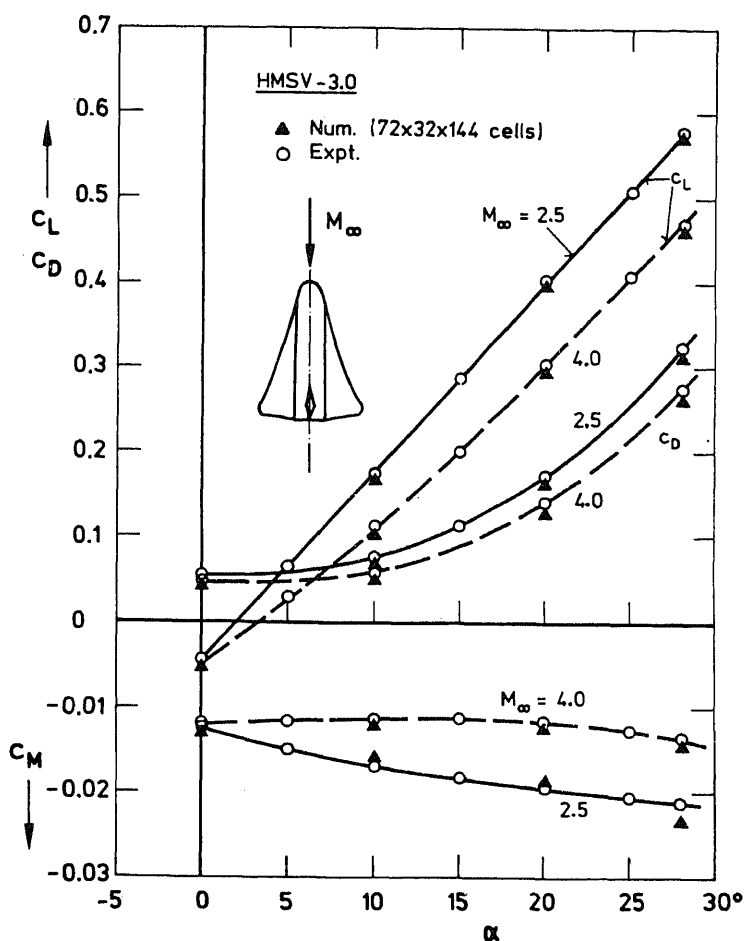
Space shuttle configurations for numerical study by using Euler and N-S equations. (a) US-Space Orbiter; (b) Space Orbits; (c) O - O grids; (d) C - O grids.



**Figure 16.** Pressure distributions and total forces and moments of the US-Space Orbiter at  $M_\infty = 0.4$  as yielded by the solution of N-S equations. (a) Cross-wise  $c_p$  distribution ( $\alpha = 21^\circ$ ; — experiment; —○— numerical); (b) forces and moment [ $M_\infty = 0.176$ ;  $Re = 9 \times 10^6$ ; ○ numerical; ● experimental (Bornemann & Surber 1978); ▽ experimental (Radespiel & Quast 1989)].

thus creating local supersonic zones on the wing upper surface. As a result, shock waves and shock boundary layer interactions are very common at the cruising condition. The transonic flow being extremely sensitive to any changes in the boundary condition, also to the formation of boundary layer thickness, it is essential to consider the effect of viscosity on the surface flow of the wing. This can be realized by basing the numerical studies on the solution of Navier–Stokes equations or else combining the solution of Euler equations with iterative correction of boundary layer displacement thickness.

In the case of the wing alone it is essential to base the numerical studies on the solution of the Navier–Stokes equations with eddy-viscosity modelling of the viscous stresses, as has been undertaken by Radespiel (1989). Two examples are considered in the paper cited, one being the transonic flow around an aerofoil (RAE-2822) at  $M_\infty = 0.73$ ,  $\alpha = 2.79^\circ$  and  $Re = 6.5 \times 10^6$ , depicting the  $c_p$ -distribution on the wing surface as well as the skin friction on the suction side. The calculated values are in excellent agreement with the experimental results. The second example concerning the flow around an ONERA-M6 wing at  $M_\infty = 0.84$ ,  $Re_\infty = 1.1 \times 10^6$  and  $\alpha = 3.06^\circ$  and  $6.06^\circ$ , will be taken up here to demonstrate the complex physics which may become involved.

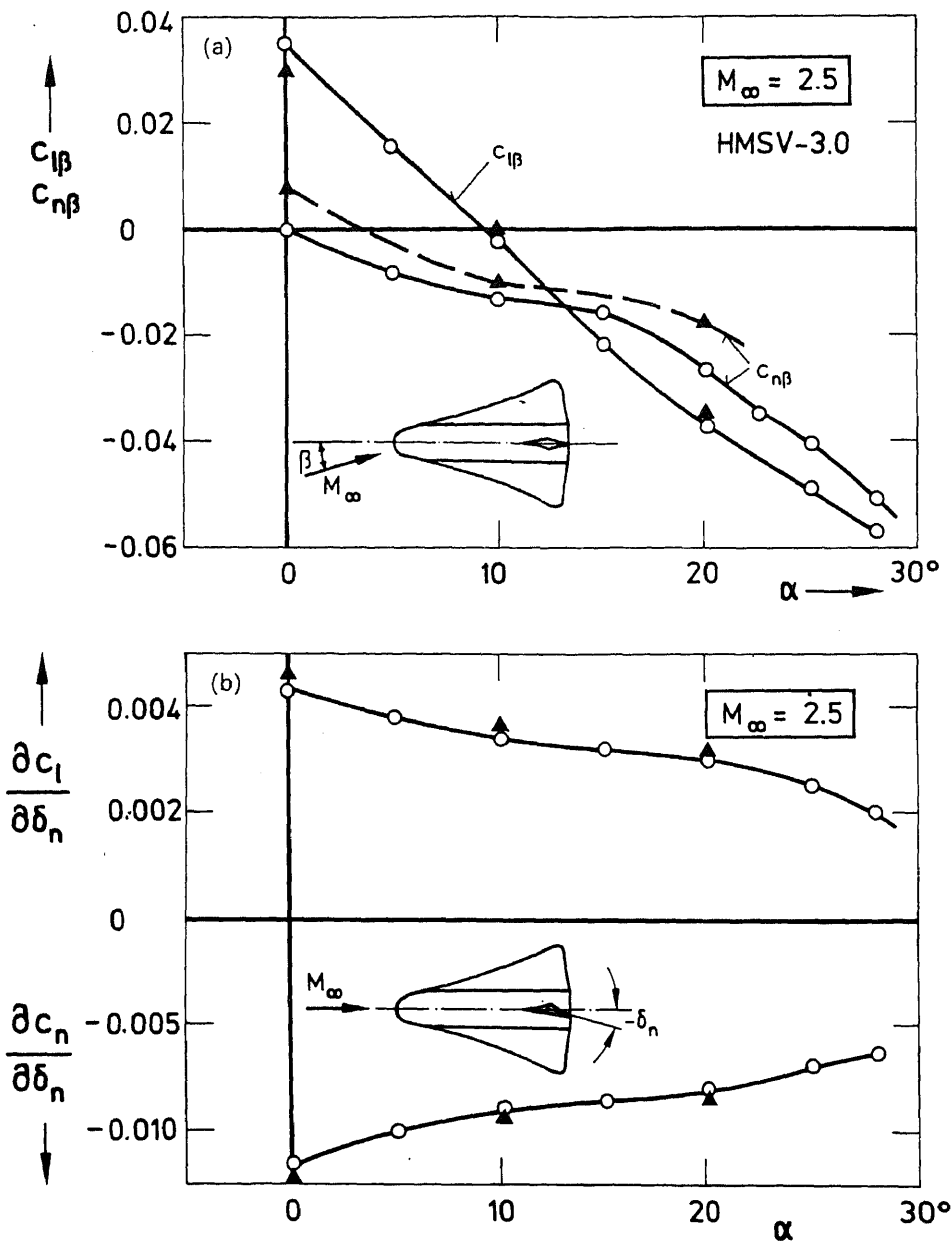


**Figure 17.** Numerical study of aerodynamic force and moment coefficients of the space orbiter Hermes by using Euler equations. [▲ (Schöne & Bidault 1991); ○ (Esch 1989).]

While for attached flow the numerical calculations performed with Baldwin–Lomax viscous modelling (Baldwin & Lomax 1978) reproduce the  $c_p$ -distribution of the wind tunnel tests very closely, as shown in figure 19 for  $\alpha = 30^\circ$ , one has to use the more complex Johnson–King (1984) modelling to have reliable results with separated flows as is evident from the results for  $\alpha = 60^\circ$ . Both the numerical schemes prove themselves to be reliable and robust, the second method being however more complex and time consuming.

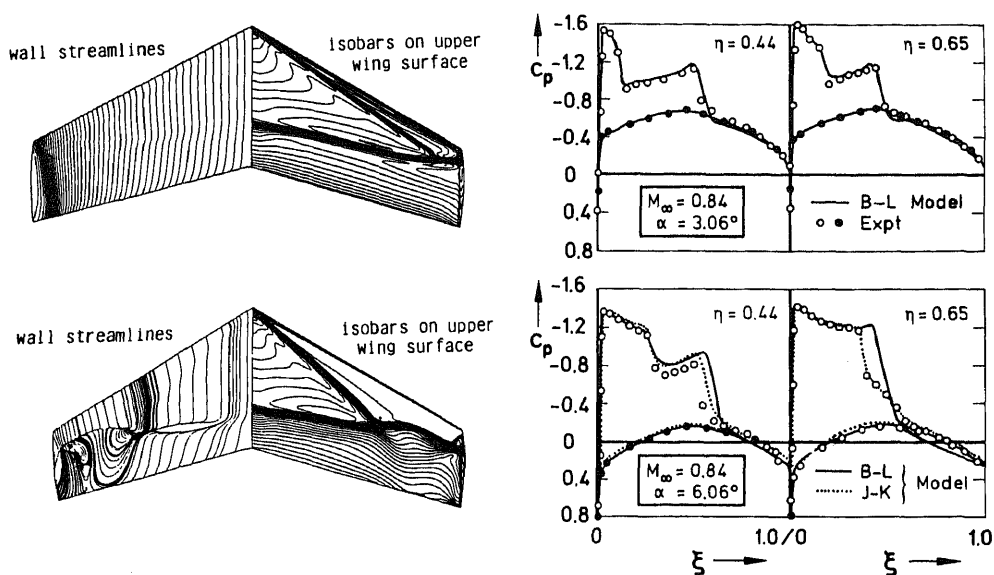
It will be now interesting to look into the interference effects of wing-fuselage configurations at high subsonic and transonic Mach numbers – at first by using the Euler equations. Extensive analysis has been undertaken for studying the flow fields of three basic configurations, which are shown in figure 20. The flow field data are used to reproduce the surface flows and isobars as well as the lift distributions and total forces. Systematic variations of geometric and aerodynamic parameters have





**Figure 18.** Analysis of the aerodynamic properties of the HERMES at the condition of side-slip (a) and rudder-deflection (b). [▲ numerical (Schöne & Bidault 1991); ○ experiment (Esch 1989).]

been undertaken for studying the nature of changes in the interference effects, especially for having higher aerodynamic efficiency of the configurations. Some essential results are depicted in figures 21 to 23. It is important to note that at transonic speeds one has to consider the viscous effects on the surface flows to get reliable results, for which some boundary-layer code has to be coupled to the numerical

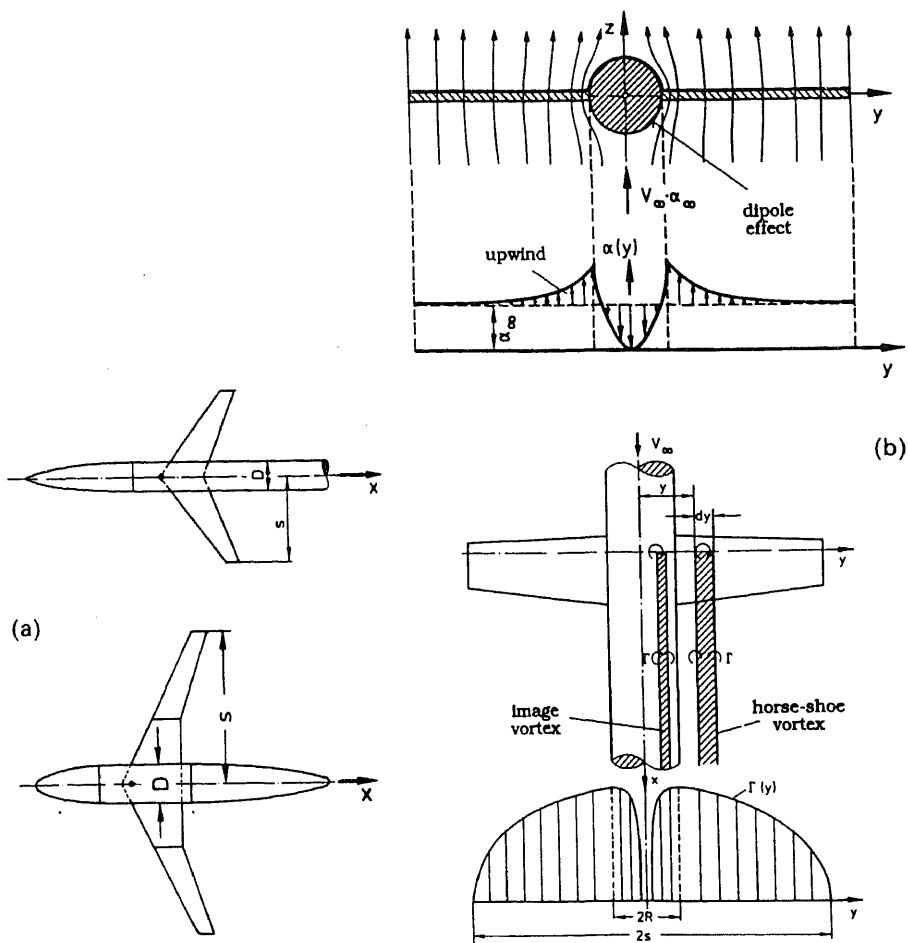


**Figure 19.** Numerical study of the flow-field of a transonic wing by using N-S equations. [B-L – Baldwin & Lomax (1978); expt. – Schmitt & Charpin (1979); J-K – Johnson & King (1984).]

scheme treating inviscid flows. For all details of the numerical methods which have been followed in the above examples one can refer to the thesis by Wichmann (1992). It is evident that numerical studies using Navier–Stokes equations can give more reliable results quite directly – they are however more time consuming and hence expensive for comprehensive studies. A few test cases have been performed (Longo 1992) yielding very good results.

The numerical studies are finally aimed at determining optimum aerodynamic configurations of complete aircraft, including wings, fuselage, engines, nacelles and pylons, especially for achieving high values of lift to drag ratios for maximizing the parameter  $M_\infty c_L/c_D$  at transonic speed regimes. Although the effect of viscosity brings in significant changes in the surface flow of a super-critical aerofoil and thus to the upper surface pressure distributions, it was decided to make the numerical calculations by using the Euler equations as a preliminary study of the global features of the flow around the complex configurations. The viscous corrections for the sensitive wing surface flow can be done by coupling a numerical code for boundary layer flows.

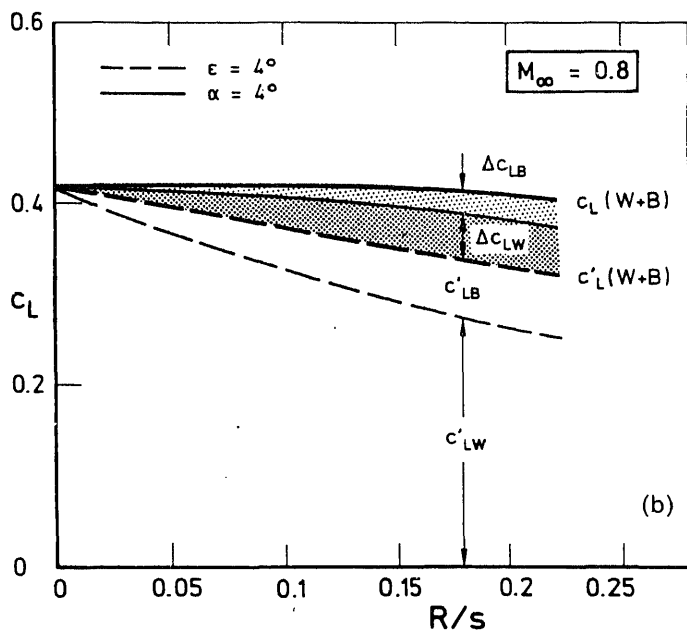
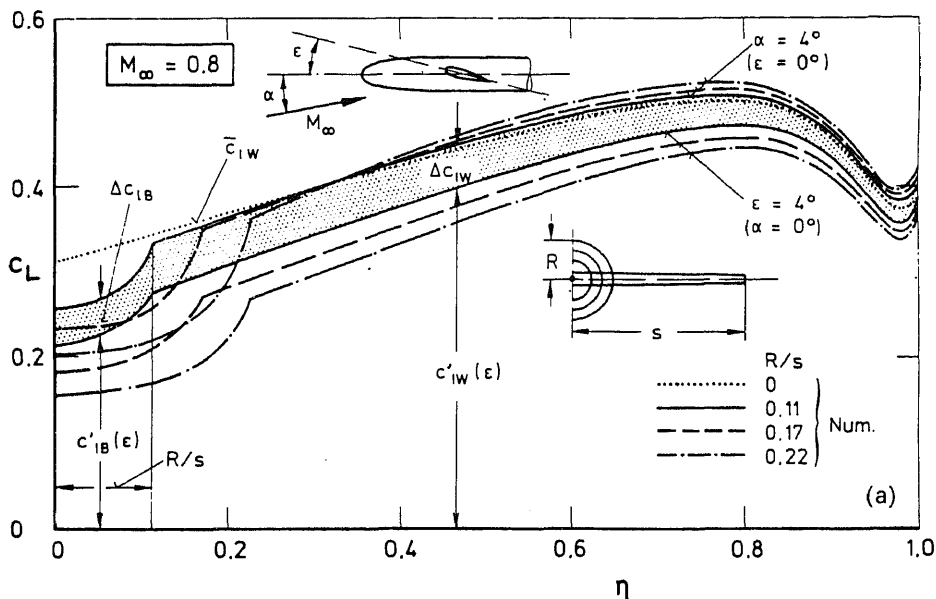
The aircraft configuration with engine, nacelle and pylon is shown in figure 24 along with the sectional marking for close study of the  $c_p$ -distributions. The computational studies also yield extensive results as surface streamlines and isobars as well as field values of pressures and velocities as isoline plots. From the surface  $c_p$ -values some plots of spanwise lift distributions are demonstrated showing the aerodynamical interference effects resulting from the engine, nacelle and pylon. For further results one may refer to Rossow & Ronzheimer (1991), Rudnik (1991) and Rossow *et al* (1992).



**Figure 20.** Wing-body combinations for numerical analysis of aerodynamic interference-effects by using Euler equations. (a) Wing-body configurations; (b) interference effects (idealized models).

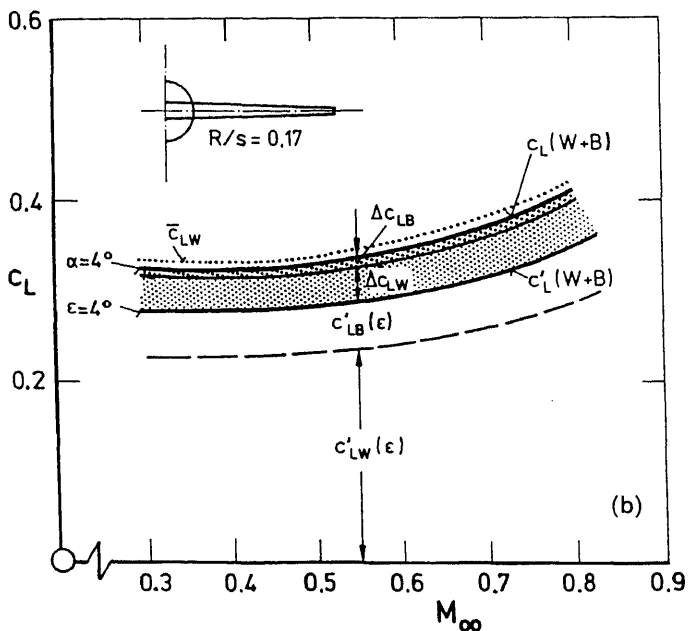
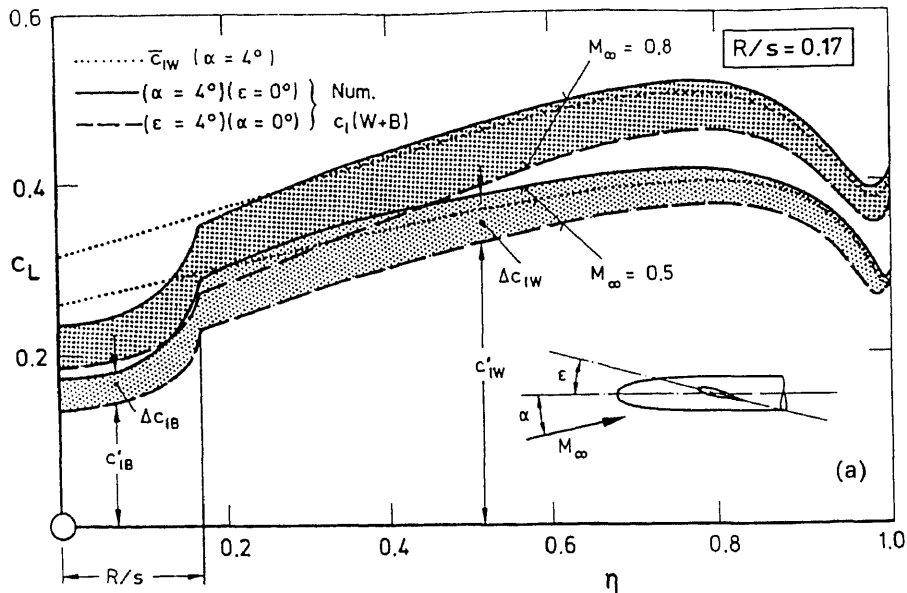
#### 4. Numerical study of flow fields around propellers and helicopter rotors

The modern aircraft-propellers and helicopter-rotors having transonic onflow conditions at the outer regions of the blades need extensive analysis for improved layout of the blade sections as well as of the blade shapes. In order to capture all the details of the flow, especially the appearance of shock waves, shock-boundary layer interactions and the vortical flow round the tips, it is essential to undertake numerical studies by using the Navier-Stokes equations. However, as a preliminary step the computational analysis was based on the Euler equations and two cases have been taken up, for which enough experimental data were available for validating the numerical results. These are a two-bladed propeller having tip advance ratio of  $\Lambda = 0.73$  with tip helical Mach number  $M_h = 0.56$  and a hovering helicopter rotor with tip Mach number of  $M_h = 0.79$ .

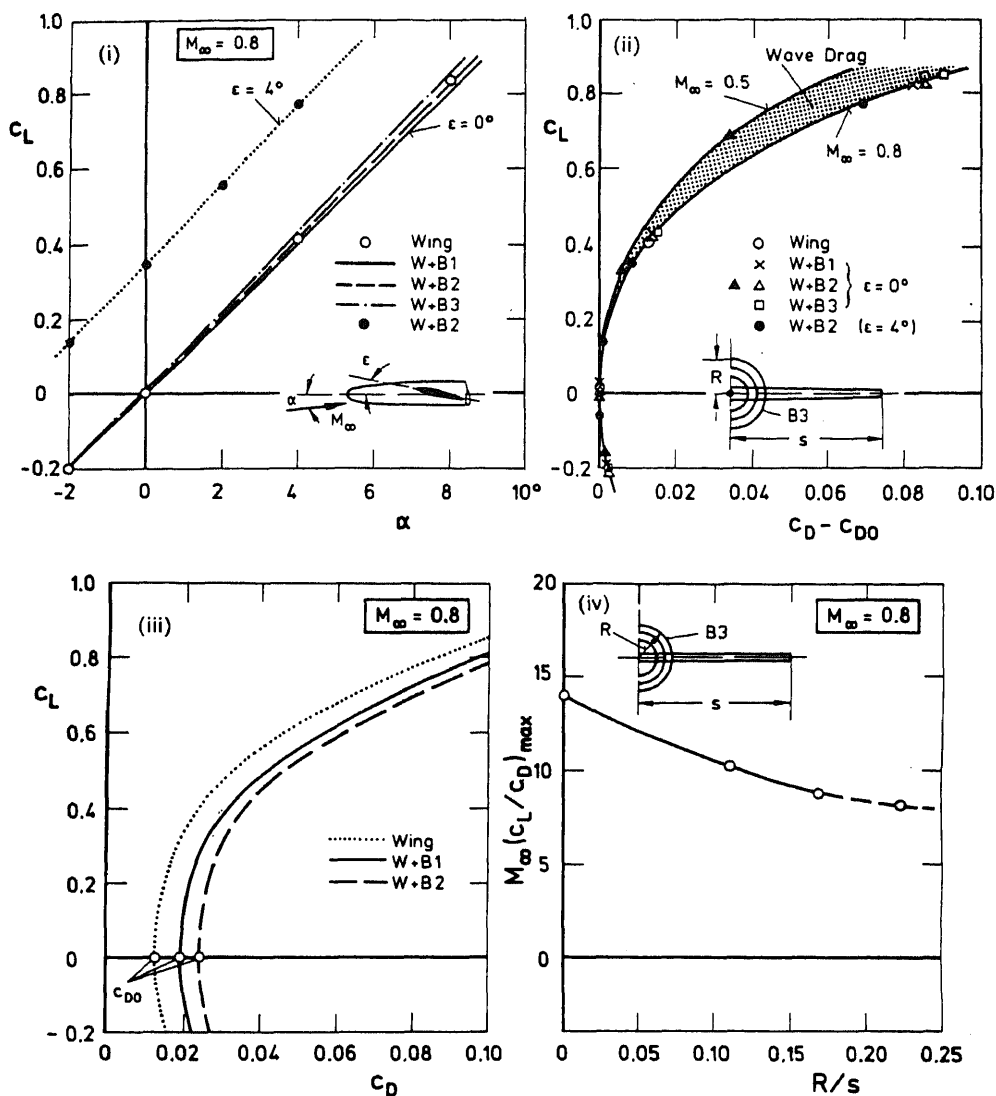


**Figure 21.** Analysis of the aerodynamic properties of wing-body combinations as function of wing-setting angle [numerical values of  $R/s$  from Wichmann (1992)] (a) and body thickness (b).

From the numerical computations using 0-0 field grids round the blades one obtains all the necessary field data for undertaking plots of blade loadings and blade surface flows, as well as of the tip vortices. The  $c_p$ -distributions on the blades are shown in figures 25 and 26 confirming very good agreement between the computed and



**Figure 22.** Effect of Mach-number on the spanwise lift-distribution (numerical values from Wichmann 1992) (a) and total-lift (b) of a wing-body combination.

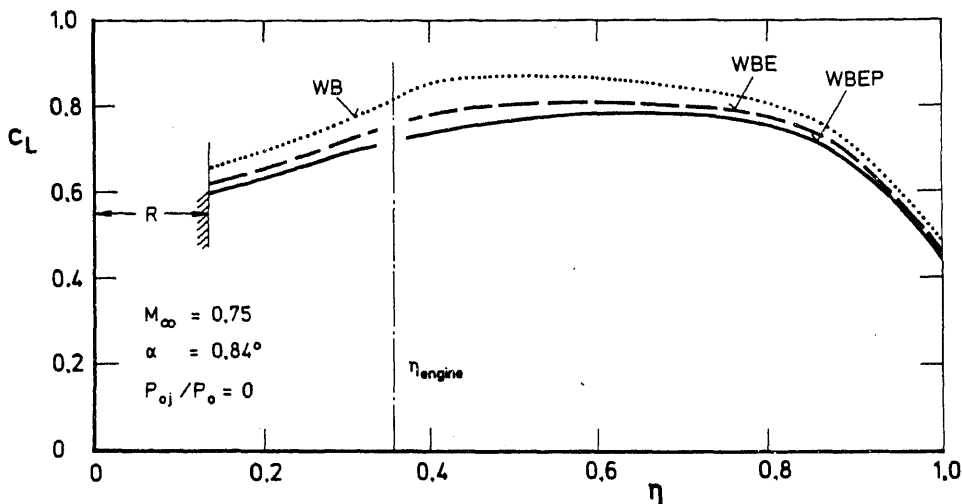
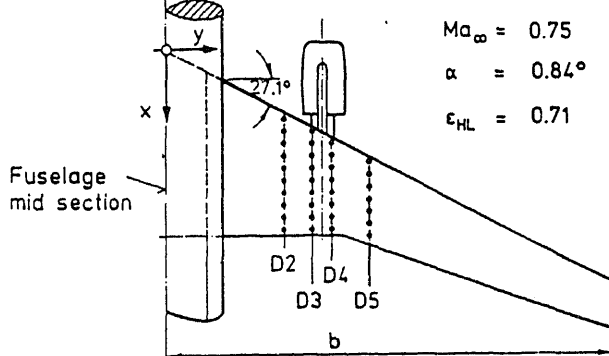


**Figure 23.** Numerical analysis of the maximum lift to drag ratio of wing-body combinations at the transonic Mach number of  $M_\infty = 0.8$ . (Numerical values in (iii) from Wichmann 1992.)

measured data. For the details of the computational method using rotating reference frame for the Euler equations one can refer to the original thesis work (Kroll 1989). The above analysis will be useful both for aerodynamics and aeroacoustics.

## 5. Numerical study of some unsteady flow fields

Some common examples of unsteady aerodynamics are oscillating and plunging motions of aerofoils and wings, both having arbitrary steady forward motion. In recent years extensive studies have been undertaken on the unsteady motions of



**Figure 24.** Numerical study of the changes in the spanwise lift distribution of a wing-body due to engine-installation.

aerofoils and wings by using numerical solutions of the Euler equations and validating the results with experimental data. Some numerical results on the unsteady forces and moment of an oscillating aerofoil are shown in figure 27 demonstrating the typical hysteresis effects.

Both with regard to aerodynamics and aeroacoustics much effort is now focused on the detailed study of unsteady flow fields of lifting helicopter rotors in forward motion. However, it seems essential to divide the numerical and experimental studies into three distinct stages:

- Flow field of a two-bladed nonlifting rotor in forward motion.
- Flow field of a two-bladed lifting rotor in forward motion.
- Flow field of a lifting multi-bladed rotor in forward motion.

The first case being much simpler than the other two will help to develop the numerical technique and some basic concepts before taking up the full problem with complex motion of the blades.

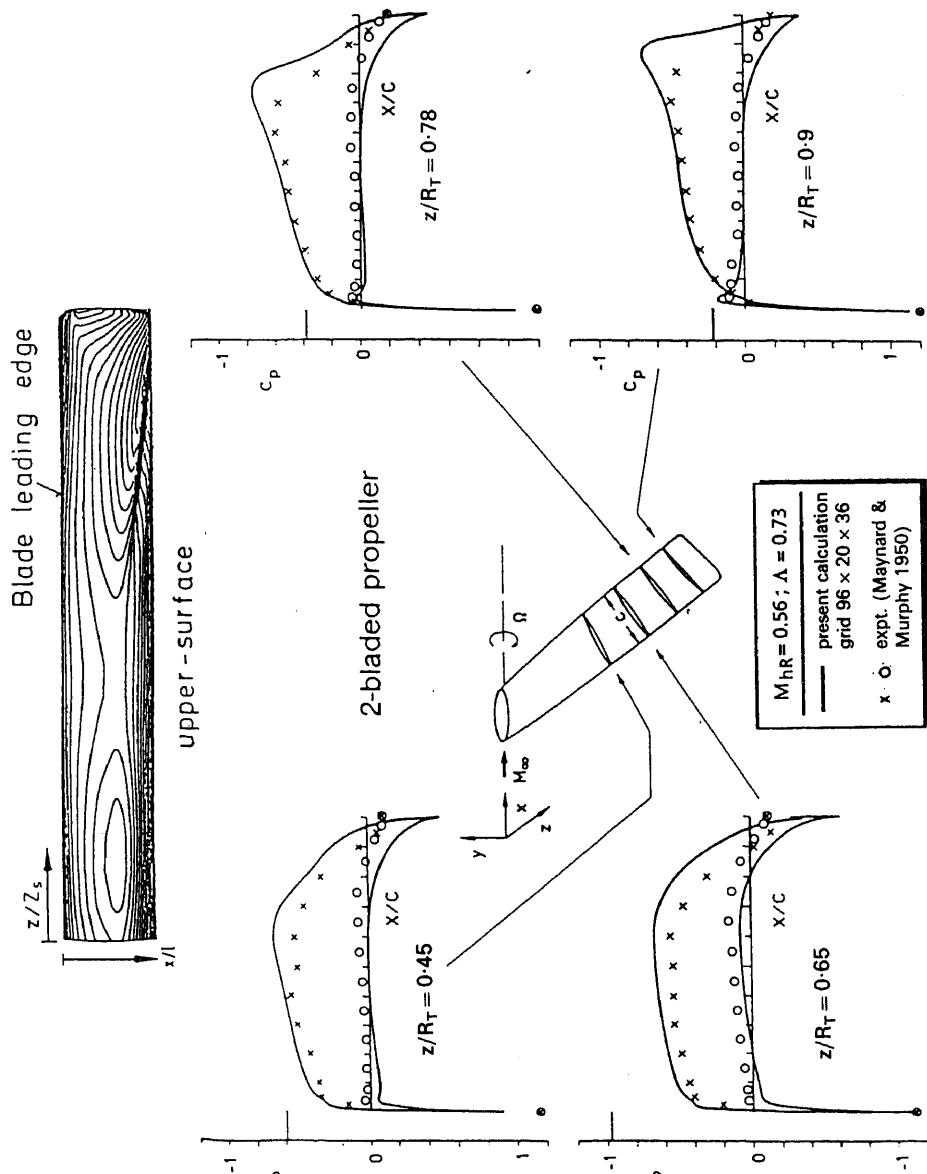
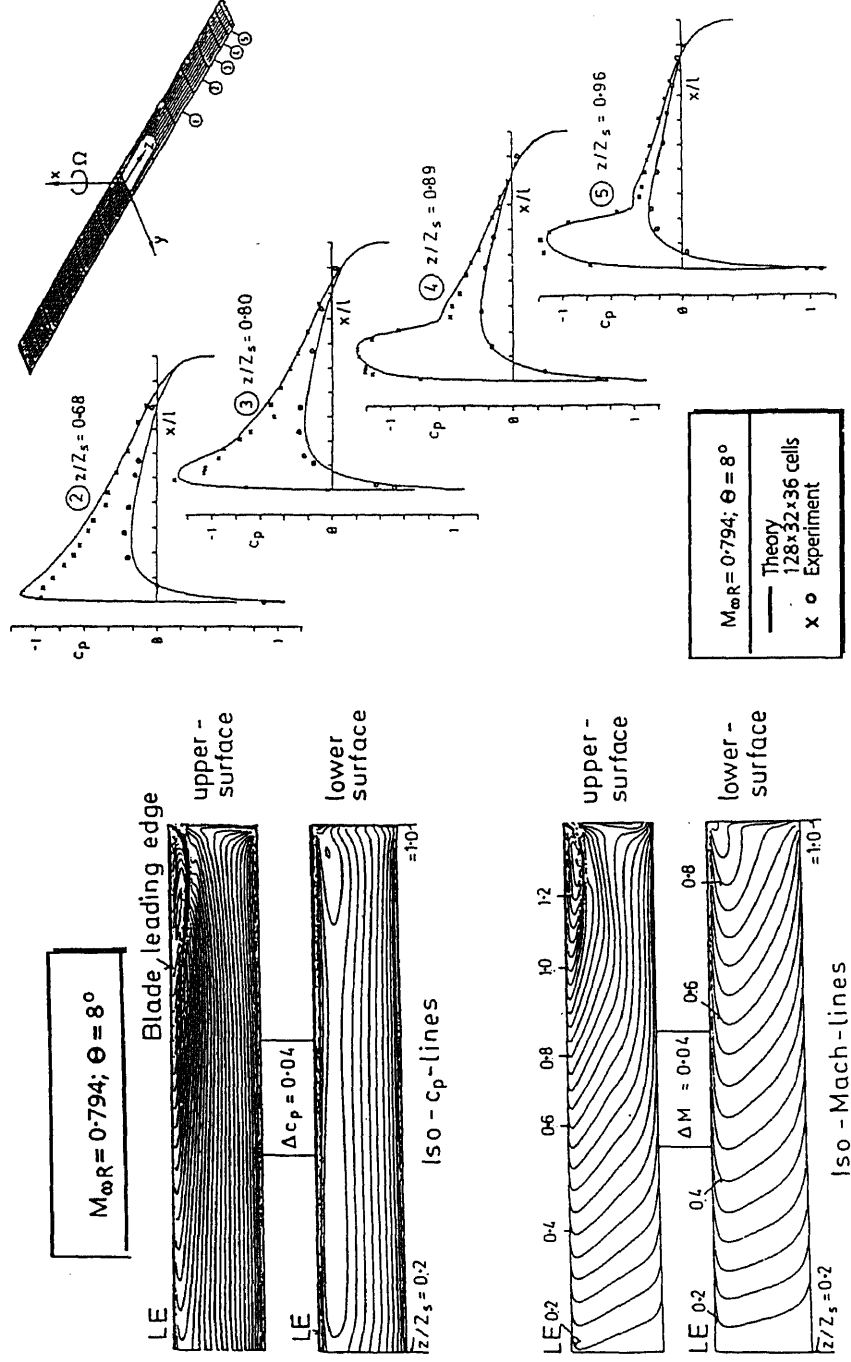
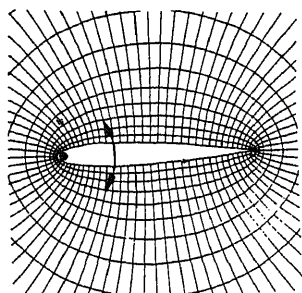


Figure 25. Numerical study of the flow field of a propeller in forward motion by using the solution of Euler equations.





**Figure 26.** Numerical analysis of the flow field of a hovering rotor by using the Euler equations. (Experimental values  $\times$  o, from Caradonna & Tung 1981.)



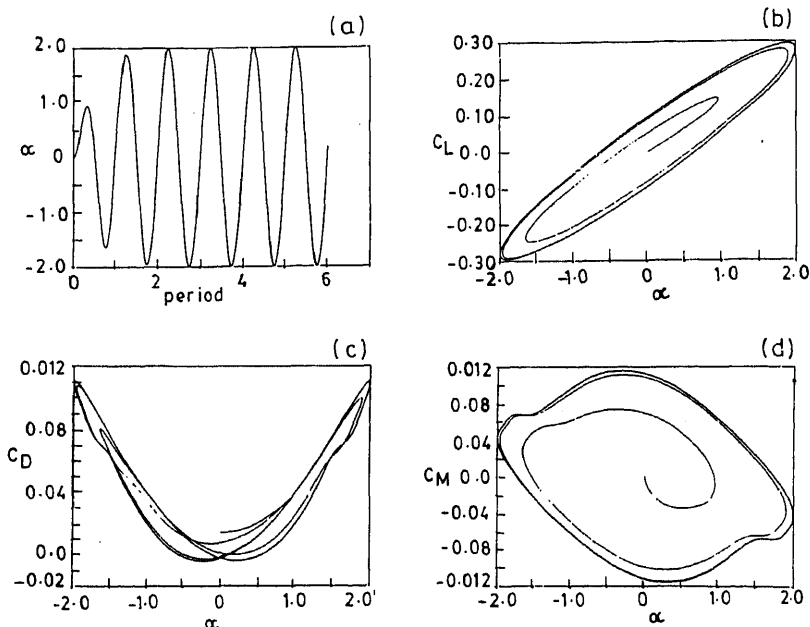
$$M_\infty = 0.755$$

$$\omega = 0.1454$$

$$k = 0.0814$$

$$\alpha = 2^\circ [1 - e^{-\sigma t}] \sin \omega t$$

Aerofoil: NACA 0012

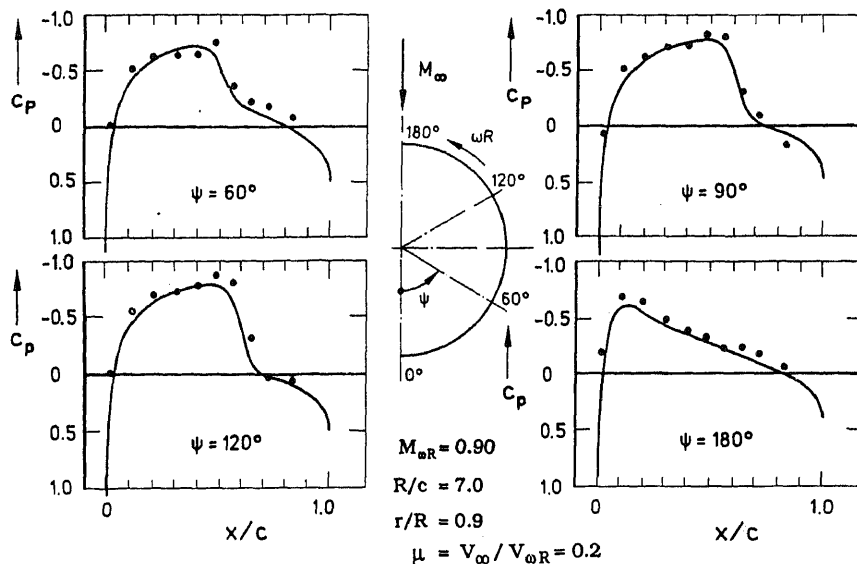


**Figure 27.** Numerical study of the unsteady flow field of an oscillating aerofoil by using the Euler equations.

Having the experimental data of  $c_p$ -distributions at the outer part of a nonlifting rotor blade, some comparisons of the numerical and experimental values at three sections with  $r/R = 0.855$  to  $0.946$  have been undertaken of which a few plots are shown in figure 28 confirming acceptable agreements. This work is being continued now as a doctorate thesis for studying the lifting cases (b) and (c).

## 6. Conclusions

The numerical methods for the solution of nonlinear partial differential equations being now well established, it has given a big impetus for undertaking detailed analysis of complex flow fields, which may arise from involved physics of the flow or due to the complexity of the geometrical configurations. Although it is preferable to base



**Figure 28.** Numerical study of the unsteady flow field of a nonlifting helicopter rotor in forward motion by using the Euler equations. [3-D Euler code O-H grids ( $65 \times 17 \times 31$ ); Expt: Caradonna *et al* (1984)].

the numerical studies on the solution of Navier–Stokes equations, they may not be highly useful if the efficiency and accuracy of the algorithms need still further effort for further improvement. Especially in case of extensive studies with systematic variations of aerodynamical and geometrical parameters it may be quite suitable to work with the numerical solutions of the Euler equations to obtain all the global features of the flow fields. The viscous effects are primarily confined in the regions of surface flows, especially where high cross-gradients come into being. While for complex flows around slender delta wings and bodies, as well as for space vehicles the numerical analysis based on the Euler equations yields quite acceptable results, one has to incorporate the effect of viscosity in the solution of the Euler equations when large span wings at supercritical onflows are concerned – and this can be usually done by coupling a boundary layer code to correct for the boundary layer displacement thickness.

Detailed analysis of the unsteady flow fields, which arise around counter-rotating propfans and helicopter rotors in forward flight have been taken up now by using the solution of Euler equations.

When further progress has been achieved with the multigrid technique, applicable also for multiblock body-fitted grids around complex configurations, direct analysis with the Navier–Stokes equations will be more in use.

### List of symbols

$a, a^*$	local and critical speed of sound;
$A$	Jacobian matrix;
$b$	span of a wing;

$c_L, c_D, c_M$	total-lift, -drag and -moment coefficients;
$c_p$	pressure coefficient;
$\bar{c}_p, \bar{c}_v$	specific heats;
$D$	dissipation operator;
$e$	internal heat energy in unit mass of the medium;
$E$	energy content in unit mass of the medium;
$F, G, H$	flux quantities in $x, y$ and $z$ directions;
$\tilde{F}, \tilde{G}, \tilde{H}$	flux quantities in $\xi, \eta$ and $\zeta$ directions;
$h, h_0$	static and total enthalpy in unit mass of the medium;
$i, j, k$	grid notations;
$J$	determinant of Jacobian matrix;
$k$	heat flow coefficient;
$l(\eta), l_0$	chord length and total length of a delta wing;
$\bar{l}$	length scale of turbulent exchanges;
$M, M_\infty$	local and onflow Mach number;
$M^*$	Mach number based on critical speed of sound;
$n$	normal to a surface;
$p, p_0$	static and total pressure;
$P_i$	source terms for grid generation;
$\tilde{q}$	heat flow;
$Q$	flux balance in elemental volume;
$r, r_0$	radial distance in the spiralling vortex and to the wing leading edge;
$R$	gas-constant;
$Re$	Reynolds number;
$s$	entropy;
$\bar{s}(\xi), \bar{s}$	local and maximum half-span of the wing;
$S$	surface area;
$S_w, S_c$	surface area of the wing and the canard;
$t$	time;
$T, T_0$	static and total temperature;
$U$	physical variables in a flow field;
$u, v, w$	velocity components in Cartesian coordinates;
$v_r, v_\theta, v_x$	velocity components in cylindrical coordinates;
$v_a$	axial velocity along the vortex core;
$v_c, v_h$	crossflow and helical velocity in the spiralling vortex;
$V, V_\infty$	local and onflow velocity;
$X^i$	physical coordinates;
$x, r, \theta$	cylindrical coordinates;
$x, y, z$	Cartesian coordinates;
$\xi, \eta, \zeta$	curvilinear coordinates;
$\alpha$	angle of incidence;
$\beta$	angle of yaw;
$\varepsilon$	geometric setting angle of the canard;
$\varphi$	sweep angle of the wing leading edge;
$\gamma$	vorticity in the flow field;
$\kappa$	ratio of specific heats;
$\rho, \rho_0$	static and stagnation medium density;
$\nu$	kinematic coefficient of viscosity;
$\mu, \mu_m, \mu_t$	coefficients of viscosity arising from molecular and turbulent exchanges;

$\sigma_v$	stress tensor due to viscosity;
$\Omega, \tilde{\Omega}$	volume of a cell element in Cartesian and curvilinear coordinate system;
$\chi$	flux tensor in the flow field.

## References

- Agrawal S, Barnett R M, Robinson B A 1990 Investigation of vortex breakdown on a delta wing using Euler and Navier-Stokes equations. *AGARD fluid dynamics panel symposium on vortex flow aerodynamics*. AGARD-CP-494, pp. 24.1-24.12
- Baldwin B S, Lomax H 1978 Thin layer approximation and algebraic model for separated turbulent flows. *AIAA paper* 78-257
- Beam R M, Warming R F 1976 Upwind second order difference schemes and application to aerodynamics. *AIAA J.* 14: 1241-1249
- Beam R M, Warming R F 1978 An implicit factored scheme for the compressible Navier-Stokes equations. *AIAA J.* 16: 385-402
- Bergmann A, Hummel D, Oelker H -Chr 1990 Vortex formation over a closed coupled canard-wing-body configuration in unsymmetrical flow. *AGARD Fluid Dynamics Panel Symposium on Vortex Flow Aerodynamics*, AGARD-CP-494, pp. 14.1-14.14
- Blazek J 1994 *Verfahren zur Beschleunigung der Lösung der Euler- und Navier/Stokes Gleichungen bei stationären Über- und Hyperschallströmungen*. Dissertation der Technische Universität, Braunschweig
- Blazek J, Kroll N, Radespiel R, Rossow C-C 1991 Upwind implicit residual smoothing method for multi-stage schemes. *AIAA paper* 91-1533
- Bober L J, Chaussee D S, Kutler P 1983 Prediction of high speed propeller flow using a three-dimensional Euler analysis. *AIAA paper* 83-0188
- Bornemann W E, Surber T E 1978 Aerodynamic design of the space shuttle orbiter. AGARD-CP-247, pp. 11.1-11.24
- Brennenstuhl U, Hummel D 1982 Vortex formation over double-delta wings. *ICAS Proc.* vol. 2, pp. 1133-1146
- Caradonna F X, Laub G H, Tung C 1984 An experimental investigation of the parallel blade-vortex interaction. NASA-TM-86005
- Caradonna F X, Tung C 1981 Experiments and analytical studies of a model helicopter rotor in hover. NACA TM-81232
- Carstens V 1990 Die Berechnung von instationären kompressiblen Strömungen um schwingende Profile mit Hilfe eines Euler-Upwind Verfahrens. DLR-Festschrift, Göttingen, 1990
- Chaussee D S 1986 High speed viscous flow calculations about complex configurations. AGARD-CP-412, pp. 29.1-29.17
- Chaussee D S, Rizk Y M, Buning P G 1984 Viscous computation of a space-shuttle flow field. *Lecture Notes in Physics - 9th Int. Conf. on Numerical Method in Fluid Dynamics* 218: 148-153
- Courant R, Friedrichs K O, Lewy H 1928 Über die partiellen Differential-gleichungen der Physik. *Math. Ann.* 100: 32-74
- Courant R, Isaacson E, Rees M 1952 On the solution of nonlinear hyperbolic differential equations by finite differences. *Commun. Pure Appl. Math.* 5: 243
- Das A 1991 Analysis of spiraling vortical flows around slender delta wings moving in an inviscid medium. *Z. Angew. Math. Mech.* 71: 465-471
- Das A, Longo J M A 1994a Numerical computation of vortical flow fields of double-delta wings moving in a compressible medium. *Z. Angew. Math. Mech.* 74: 475-486
- Das A, Longo J M A 1994b Numerical analysis of the vortical flow field around a delta-shaped wing-canard configuration moving in a compressible medium (to be published in *AIAA J.*)
- Deese J E, Agarwal R K 1987 Navier-Stokes calculations of transonic viscous flow about a wing body configuration. *AIAA paper* 87-1200
- Deese J E, Agarwal R K 1988 Euler/Navier-Stokes calculations of transonic flow past fixed- and rotary-wing aircraft configurations. *Transonic Symposium Proceedings* vol. 1, part 2, NASA-LARC, Hampton, pp. 521-545

- Degani D, Schiff L B 1983 Computation of supersonic viscous flows around pointed bodies at large incidence. *AIAA paper* 83-0034
- Drouge G 1988 The international vortex flow experiment for computer code validation. *ICAS-Proceedings 1988*, vol. 1, pp. 25-41
- Eberle A 1987 Characteristic flux averaging approach to the solution of Euler's equation. *VKI Lecture Series* 1987-04
- Elsenaar A, Hoijemakers H W M 1990 An experimental study of the flow over a sharp-edged delta wing at subsonic and transonic speeds. *AGARD FDP Symposium on Vortex Flow Aerodynamics*. AGARD-CP-494, pp. 15.1-15.19
- Eriksson L E 1982 Generation of boundary-conforming grids around wing-body configurations using transfinite interpolation. *AIAA J.* 20: 1313-1320
- Eriksson L, Rizzi A 1983 Computation of vortex flow around a canard-delta wing combination. *Proc. of 5th GAMM Conference on Numerical Methods in Fluid Mechanics* pp. 65-80
- Esch H 1989 Force measurements on four 1:100 HERMES shape-derivatives at subsonic and supersonic Mach numbers. *DLR-IB* 39113-89-C-21, 1989
- Findling A, Herrmann U 1991 Development of an efficient grid generation. *Proc. 3rd Int. Conf. on Numerical Grid Generation in Comput. Fluid Dynamics and Related Fields*. Barcelona, Spain, pp. 781-792
- Goddard J L, Jacquotte O P, Gisquet D 1991 Analyse detaillee de l'interaction voilure-nacelle d'un avion de transport civil. *AGARD Symposium on Aerodynamic Engine/Airframe Integration*, Texas
- Hartwich P M, Hsu C H, Luckring J M, Liu C H 1988 Aerodynamic application of an efficient incompressible Navier-Stokes solver. *ICAS Proc.* vol. 2, pp 1417-1427
- Hilgenstock A 1990 Ein Beitrag zur numerischen Simulation der Transsonischen Strömung um einen Deltaflügel durch Lösung der Navier-Stokes'schen Bewegungsgleichungen. *DLR-FB* 90-13
- Hilgenstock A, Vollmers H 1990 On the simulation of compressible turbulent flows past delta wings, delta wing-body and delta wing-canard. *AGARD FDP Symposium on Vortex Flow Aerodynamics*, AGARD-CP-494, pp. 7.1-7.14
- Hummel D 1988 Documentation of separated flows for computational fluid dynamics validation. AGARD-CP-437, vol. 2, pp. 15.1-15.24
- Jameson A 1985 Multigrid algorithms for compressible flow calculations. *MAE Report* 1743, Princeton University
- Jameson A, Schmidt W, Turkel E 1981 Numerical simulation of the Euler equations by finite volume method using Runge-Kutta time stepping schemes. *AIAA paper* 81-1259
- Johnson D A, King L S 1984 A new turbulence closure model for boundary layer flows with strong adverse pressure gradients and separation. *AIAA paper* 84-0175
- Krause E, Liu C H 1989 Numerical studies of incompressible flow around delta and double-delta wings. *Z. Flugwiss.* 13: 291-301
- Kroll N 1989 Berechnung von Strömungsfeldern um Propeller und Rotoren im Schwebeflug durch die Lösung der Euler-Gleichungen. Dissertation Technische Universität Braunschweig, *DLR-FB* 89-37
- Kroll N, Gaitonde D, Aftosmis M 1991 A systematic comparative study of several high resolution schemes for complex problems in high speed flows. *AIAA paper* 91-0636
- Kroll N, Jain R 1987 Solution of two-dimensional Euler equations - experience with a finite volume code. *DFVLR-FB* 87-41
- Kroll N, Radespiel R, Rossow C -C 1989 Experiences with explicit time-stepping schemes for supersonic flow fields. *8th GAMM Conf. on Numerical Methods in Fluid Dynamics* (Delft: Vieweg)
- Kroll N, Rossow C -C 1990 A high resolution cell vertex TVD scheme for the solution of the two- and three-dimensional Euler equations. *12th Int. Conf. on Num. Methods in fluid dynamics. Lecture Notes in Physics* 371:
- Kumar A, Das A 1988 Computation of vortex flow on a delta wing at transonic speed. *IUTAM Symposium Transsonicum III* pp. 317-328
- Lambourne N C 1982 Compendium of unsteady aerodynamics measurements, AGARD-R-702
- Lax P D, Wendroff B 1966 System of conservation laws. *Commun. Pure Appl. Math.* 13: 217-237
- Lin C Q, Pahlke K 1991 Numerical solution of Euler equations for airfoils in arbitrary unsteady motion. *DLR-FB* 129-91/17

- Longo J M A 1988 Research on three different Euler schemes applied to a delta wing with vortical flows. *DGLR STAB Symposium*
- Longo J M A 1992a *Untersuchungen der Umströmung von Deltaflügel-Konfigurationen mit und ohne Canard durch numerische Lösung der Euler-Gleichungen*. Dissertation Technische Universität Braunschweig
- Longo J M A 1992b Viscous transonic flow simulation around a transport aircraft configuration. *DGLR Jahrestagung*
- Longo J M A, Das A 1990 Numerical simulation of vortical flows over close-coupled canard-wing configuration. *AIAA* 90-3003
- MacCormack R 1982 A numerical method for solving the equations of compressible viscous flow. *AIAA J.* 20: 1275-1281
- MacCormack R 1985 Current status of numerical solutions of the Navier-Stokes equations. *AIAA* paper 85-0032
- Maynard J D, Murphy M P 1950 Pressure distributions on the blade sections of the NACA 10(3) (066)-033 propeller under operating conditions. *NACA RM L9L12*
- Murman E M, Rizzi A 1986 Application of Euler equations to sharp edge delta wing with leading edge vortices. *AGARD Conf. Proc. No. 412* pp. 15-1 to 15-13
- Newsome R W, Kandil O A 1987 Vortical flow aerodynamics - Physical aspects and numerical simulation. *AIAA* paper 87-0205
- Nixon D (ed.) 1989 Unsteady transonic aerodynamics. *AIAA, Progress in Astronautics and Aeronautics*, vol. 120
- Oelker H -Chr 1990 *Aerodynamische Untersuchungen an kurzgekoppelten Entenkonfigurationen bei symmetrischer Anströmung*. Dissertation Technische Universität Braunschweig, ZLR Forsc. Bericht der Technische Universität Braunschweig 90-01
- Pahlke K, Kroll N 1991 Grid generation for multi-blades and calculations with a 3-D Euler-code for helicopter rotors in hover. *DLR-IB* 129-91/4
- Pan D, Lomax H 1986 A new approximate LU factorization scheme for the Reynold's-averaged Navier-Stokes equations. *AIAA* paper 86-0337
- Pfützner M, Weiland C 1987 Three-dimensional Euler solution for hypersonic Mach numbers. *AGARD-CP* 429
- Pulliam T H, Steger J L 1980 Implicit finite difference simulation of three-dimensional compressible flow. *AIAA J.* 18: 159-167
- Radespiel R 1988 Grid generation around transport aircraft configurations using a multiblock structured computational domain. In *AGARDograph No. 309* pp. 139-153
- Radespiel R 1989 A cell-vertex multigrid method for the Navier-Stokes equations. *NASA-TM* 101557
- Radespiel R, Kroll N 1985 Progress in the development of an efficient finite volume code for the three dimensional Euler equations. *DFVLR-FB* 85-31
- Radespiel R, Kroll N 1990 A multigrid scheme with semicoarsening for accurate computation of viscous flows. *12th Int. Conf. on Numerical Methods in Fluid Dynamics. Lecture Notes in Physics*. 371:
- Radespiel R, Quast A 1989 Kraftmessungen, Druckverteilungen und Strömungssichtbarmachung am Raumgleitermodell FALKE im Niedergeschwindigkeitsbereich. *DLR IB-129-89/37*
- Radespiel R, Rossow C -C, Swanson R C 1990 An efficient cell-vertex multigrid scheme for the three-dimensional Navier-Stokes equations. *AIAA J.* 28: 1464-1472
- Radespiel R, Swanson R C 1991 Progress with multigrid schemes for hypersonic flow problems. *NASA CR 189579 ICASE Report No. 91-89*
- Raj P, Sikora J S, Keen J M 1988 Free vortex flow simulation using a three-dimensional Euler aerodynamics method. *ICAS Proc.* vol. 1, pp. 604-617
- Redeker G, Müller R, Ashill P R, Elsenaar A, Schmitt V 1987 Experiments on the DFVLR-F4 wing-body configuration in several European wind tunnels. *AGARD-CP* 429, pp. 2.1-2.15
- Rizetta D P, Shang J S 1987 Numerical simulation of leading edge vortex flows. *AIAA J.* vol. 24, pp. 237-245
- Rizk Y M, Chaussee D S, McRae D S 1981 Computation of hypersonic flows around three-dimensional bodies at high angles of attack. *AIAA* paper 81-1261
- Roe P L 1981 Approximate Riemann solvers, parameter vectors and difference schemes. *J. Comput. Phys.* 43: 357-372
- Rossow C -C 1987 Comparison of cell centered and cell vertex finite volume schemes. In *Proc. 7th GAMM Conference, Louvain-La-Neuve/Belgium*

- Rossow C -C 1989 *Berechnung von Strömungsfeldern durch Lösung der Euler-Gleichungen mit einer erweiterten Finite-Volumen Diskretisierungsmethode*. Dissertation der Technische Universität Braunschweig, DLR-FB 89-38
- Rossow C -C 1991a Flux balance splitting: A second order accurate cell vertex upwind scheme. *9th GAMM Conference on Numerical Methods in Fluid Mechanics*, Lausanne/Schweiz
- Rossow C -C 1991b Efficient computation of inviscid flow fields around complex configurations using a multiblock multigrid method. *5th Copper Mountain Conf. of Multigrid Methods*
- Rossow C -C, Goodard J -L, Hoheisel H, Schmidt V 1992 Investigation of propulsion integration interference effects on a transport aircraft configuration. *28th Joint Propulsion Conference*, Nashville, TN, AIAA paper 92-3097
- Rossow C -C, Ronzheimer A 1991a Investigation of interference phenomena of modern wing-mounted high-bypass-ratio engines by the solution of the Euler-equations. *AGARD Symposium on Engine/Airframe Integration*, AGARD-CP-98, pp. 5.1-5.11
- Rossow C -C, Ronzheimer A 1991b Multiblock grid generation around wing-body-engine-pylon configurations. *Numerical Grid Generation in Computational Fluid Dynamics and Related Fields* pp. 357-368
- Rudnik R 1991 Erweiterung eines dreidimensionalen Euler-Verfahrens zur Berechnung des Strömungsfeldes um Nebenstromtriebwerke mit Fan- und Kernstrahl. *Z. Flugwiss.* 15(5):
- Scherr S, Das A 1988 Basic analysis of the flow fields of slender wings using the Euler equations. *ICAS Proc.* vol. 2, pp. 1428-1436
- Schmitt V, Charpin F 1979 Pressure distribution on the ONERA M6 wing at transonic Mach numbers. AGARD-AR-138
- Schöne J, Bidault J 1991 Calculation of inviscid flow around a reentry configuration at supersonic speed. AIAA paper 91-0391
- Schöne J, Streit Th, Kröll N 1990 Aerodynamische Berechnungen für eine HERMES-Studienkonfiguration bei Überschallanströmung durch Lösung der Euler Gleichungen. DGLR-Jahrestagung 1990, 1.-4. Oktober 1990, Friedrichshafen. Jahrbuch 1990 der DGLR
- Schöne J, Streit Th, Kröll N 1991 Steps towards an efficient and accurate method solving the Euler equations around a re-entry configuration at super- and hypersonic speed. *Proceedings 1st European Symposium on Aerothermodynamics for Space Vehicles*, Noordwijk
- Schwarz W 1986 Elliptic grid generation system for three-dimensional configurations using Poisson's equation. *Proc. 1st Int. Conf. on Numerical Grid Generation in Comput. Fluid Mech.*, (Pineridge Press) pp. 341-352
- Sonar Th 1989 Grid generation using elliptic differential equations. DFLVR-FB 89-15
- Sonar Th, Radespiel R 1986 Geometric modelling of complex aerodynamic surfaces and three-dimensional grid generation. In: "Numerical Grid Generation in CFD", *Proc. 2nd Int. Conf. Miami* (Pineridge Press)
- Steger J, Warming R F 1976 Flux vector splitting of inviscid gasdynamic equations with application to finite difference methods. NASA TM-78605
- Swanson R C, Radespiel R 1991 Cell centered and cell vertex multigrid schemes for the Navier-Stokes equations. *AIAA J.* 29: 697-703
- Thompson J F, Warsi Z U H, Mastin C W 1985 *Numerical grid generation. Foundation and applications* (Amsterdam: North Holland)
- Van Leer B 1985 Upwind difference methods for aerodynamic problems governed by the Euler equations. *Lect. Appl. Math.* 23: 327-336
- Van Leer B, Thomas J L, Roe P L, Newsome R W 1987 A comparison of numerical flux formulas for the Euler and Navier-Stokes equations. AIAA paper 87-1104
- Volpe G, Jameson A 1988 An efficient method for computing transonic and supersonic flows about aircrafts. ICAS-88-4.8.3, pp. 1224-1235
- Whitfield D L, Janus D M 1984 Three-dimensional unsteady Euler equations solution using flux vector splitting. AIAA paper 84-1552
- Whitfield D L et al 1987 Three-dimensional unsteady Euler solutions for propfans and counterrotating propfans in transonic flow. AIAA paper 87-1197
- Wichmann G 1992 *Untersuchungen zur Flügel-Rumpf-Interferenz durch Anwendung eines Eulerverfahrens für kompressible Strömungen*. Dissert, Technische Universität Braunschweig





# On the boundary-layer control through momentum injection: Studies with applications\*

V J MODI<sup>1</sup> and T YOKOMIZO<sup>2</sup>

<sup>1</sup>Department of Mechanical Engineering, The University of British Columbia, Vancouver BC, Canada V6T 1Z4

<sup>2</sup>Department of Mechanical Engineering, Kanto Gakuin University, Mutsuura, Kanazawa, Yokohama, Japan 236

**Abstract.** The concept of moving surface boundary-layer control, as applied to a Joukowski airfoil, is investigated through a planned experimental programme complemented by numerical studies. The moving surface was provided by rotating cylinders located at the leading edge and/or trailing edge as well as top surface of the airfoil. Results suggest that the concept is quite promising, leading to a substantial increase in lift and a delay in stall. Depending on the performance desired, appropriate combinations of cylinder geometry, location and speed can be selected to obtain favourable results over a wide range of angle of attack. Next, effectiveness of the concept in reducing drag of bluff bodies such as a two-dimensional flat plate at large angles of attack, rectangular prisms and three-dimensional models of trucks is assessed through an extensive wind tunnel test-programme. Results show that injection of momentum through moving surfaces, achieved here by introduction of bearing-mounted, motor-driven, hollow cylinders, can significantly delay separation of the boundary-layer and reduce the pressure drag. The momentum injection procedure also proved effective in arresting wind-induced vortex resonance and galloping type of instabilities. A flow visualization study, conducted in a closed-circuit water tunnel using slit lighting and polyvinyl chloride tracer particles, adds to the wind-tunnel and numerical investigations. It shows, rather dramatically, the effectiveness of the moving surface boundary-layer control (MSBC).

**Keywords.** Boundary-layer control; bluff body aerodynamics; drag reduction.

## 1. Introduction

Ever since the introduction of the boundary-layer concept by Prandtl, there has been a constant challenge faced by scientists and engineers to minimize its adverse effects

---

\*The Sabita Chaudhury Memorial Lecture

and control it to advantage. Methods such as suction, blowing, vortex generators, turbulence promoters etc. have been investigated at length and employed in practice with varying degrees of success. A vast body of literature accumulated over years has been reviewed rather effectively by several authors including Goldstein (1938), Lachmann (1961), Rosenhead (1966), Schlichting (1968), Chang (1970) and others. However, the use of moving wall for boundary-layer control has received relatively little attention. This is indeed surprising as the associate committee on aerodynamics, appointed by the National Research Council, specifically recommended more attention in this area (National Research Council 1966).

Irrespective of the method used, the main objective of a control procedure is to prevent, or at least delay, the separation of the boundary-layer from the surface. A moving surface attempts to accomplish this in two ways: it prevents the initial growth of boundary-layers by minimizing relative motion between the surface and the free stream; and it injects momentum into the existing boundary-layer.

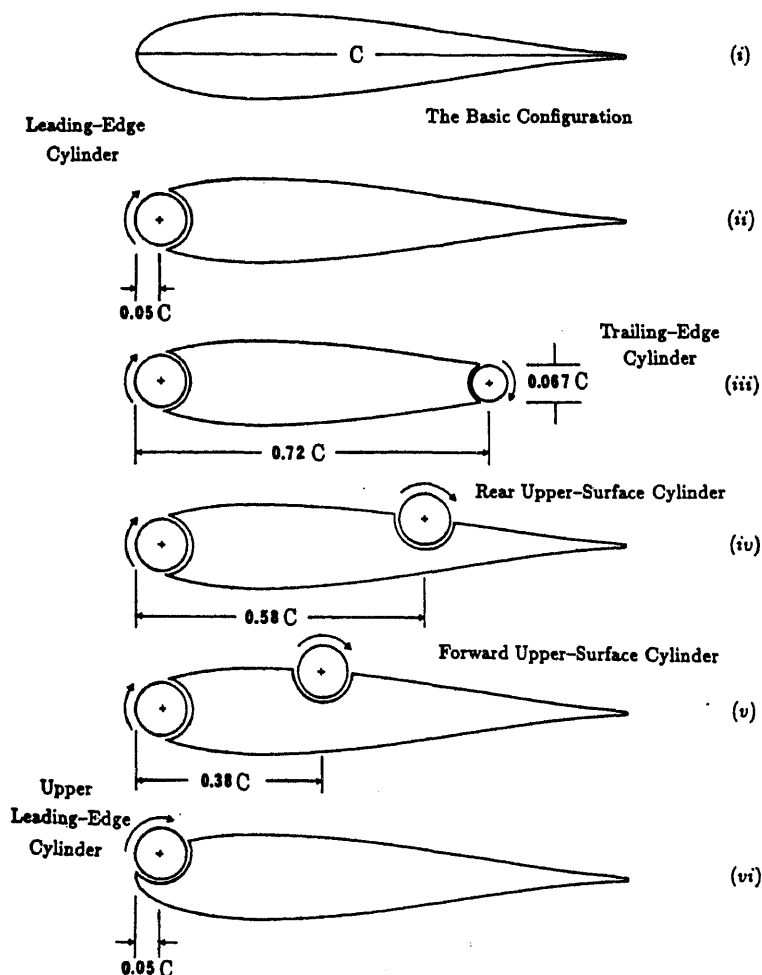


Figure 1. Various rotating-cylinder configurations studies with the Joukowski airfoil model.

Newton was probably the first one to observe the effect of moving wall boundary-layer control on the trajectory of a spinning ball (Thwaites 1960, p. 215) without any appreciation as to the basis of the effect. Almost 200 years later Magnus (1853) studied lift generated by circulation and utilized the effect to construct a ship with a vertical rotating cylinder replacing the sail. Swanson (1961) and Iverson (1972) have presented excellent reviews of literature on the Magnus effect. As early as in 1910, Prandtl (Betz 1961) himself demonstrated his "ship of zero resistance" through flow around two counter-rotating cylinders, while Flettner (1925) applied the principle to ship propulsion in 1924 when he fitted large vertical rotating cylinders on the deck of the *Buchau*. A little later, in 1934, Goldstein (1938) illustrated the principle of boundary-layer control using a rotating cylinder at the leading edge of a flat plate. However, the most practical application of the moving wall for boundary-layer control was demonstrated by Favre (1938). Using an airfoil with the upper surface formed by a belt moving over two rollers, he was able to delay separation until the angle of attack ( $\alpha$ ) reached  $55^\circ$  where the maximum lift coefficient of 3.5 was realized.

Efforts so far, though useful to an extent, were generally aimed at specific configurations and were scattered and lacked approach to the problem at a fundamental level in an organized fashion. From this point of view, the contribution by Modi and coworkers to the field is significant (Modi *et al* 1979–81, 1987a, 1987b, pp. 225–30, 1988, pp. 63–71, 1990a; Modi & Mokhtarian 1985; Mokhtarian & Modi 1984, pp. 167–75, 1986, pp. 322–30, 1988). They have studied in a comprehensive manner, the application of the moving surface boundary-layer control (MSBC) with reference to two-dimensional Joukowski airfoils having one or more cylinders acting as momentum injecting elements (figure 1). The wind tunnel results were complemented by numerical as well as flow visualization studies (Modi 1991; Modi & Yokomizo 1992, pp. 270–4). As can be expected the amount of information obtained is literally enormous. However, for conciseness, only some salient aspects of their investigations and typical results useful in establishing trends are recorded here.

## 2. MSBC as applied to two-dimensional airfoils

### 2.1 Wind tunnel test-program

The wind-tunnel model, a symmetrical Joukowski airfoil of 15% maximum thickness to chord ratio, approximately 0.38 m along the chord and 0.68 m long, spanned the tunnel test-section,  $0.91 \times 0.68 \times 2.6$  m, to create essentially a two-dimensional condition. The model was provided with pressure taps, suitably distributed over the circumference, to yield detailed information concerning the surface loading. It was supported by an aerolab six-component strain gauge balance and tested in a low-speed, low-turbulence return-type wind tunnel where the airspeed can be varied from 1–50 m/s with a turbulence level of less than 0.1%. A Betz micromanometer with an accuracy of 0.2 mm of water was used to measure the pressure differential across the contraction section of 7:1 ratio. The rectangular test-section ( $0.91 \times 0.68$  m) is provided with  $45^\circ$  corner fillets that vary from  $15.25 \times 15.25$  to  $12 \times 12$  cm to partly compensate for the boundary-layer growth. The spatial variation of velocity in the test-section is less than 0.25%.

The rotating cylinders were supported by high-speed bearings housed in the brackets at either end of the model. They were driven by 1/4 hp, 3.8 A variable speed

motors, located outside the tunnel, through standard couplings. The configurations tested include the leading-edge, trailing-edge, forward upper-surface, rear upper-surface, upper-surface, and upper leading-edge cylinders. The model was provided with a total of 44 pressure taps, distributed over the circumference, to yield detailed information about the surface loading. However, once a section of the model was removed to accommodate a cylinder, the pressure taps in that section were lost. Although the pressure information over the small region represented by the upper-surface cylinder is not of particular significance, the corresponding data at the leading edge of the airfoil are crucial since it represents a high-suction region. Its measurement presented a challenging task. Locating pressure taps on the surface of the cylinder, typically rotating in the range of 2000–8000 rpm offers considerable practical difficulty. The problem was resolved by measuring the pressure in the immediate vicinity of the cylinder rather than on the surface itself.

This was achieved in the case of the leading-edge cylinder by keeping the pressure taps stationary while the cylinder rotated. By locating the tap in a narrow ring, the width of which represented only a very small fraction of that of the cylinder, it was possible to ensure the continuity of flow over the entire surface and to obtain an estimate of the surrounding pressure. The leading-edge cylinder was provided with grooves to house the “pressure rings” which maintained the cylinder surface uniform.

The tests were conducted over an extended range of angles of attack and cylinder rotational speeds, corresponding to  $Uc/U = 0, 1, 2, 3, 4$  at a Reynolds number of  $4.62 \times 10^4$ . Here  $Uc$  and  $U$  correspond to the cylinder surface and free stream velocities, respectively. The choice of the Reynolds number in this case was dictated by vibration problems with multicylinder configurations operating at high rotational speeds (around 8000 rpm for  $Uc/U = 4$ ). The pressure plots were integrated in each case to obtain the lift coefficient. The lift was also measured independently using an aerolab six-component strain gauge balance to assess the two-dimensional character of the flow.

## 2.2 Results and discussion

The relatively large angles of attack used in the experiments result in a considerable blockage of the wind-tunnel test-section, from 21% at  $\alpha = 30^\circ$  to 30% at  $\alpha = 45^\circ$ . The wall confinement leads to an increase in local wind speed at the location of the model, thus resulting in an increase in aerodynamic forces. Several approximate correction procedures have been reported in the literature to account for this effect. However, these procedures are mostly applicable to streamlined bodies with attached flow. A satisfactory procedure applicable to a bluff body offering a large blockage in a flow with separating shear layers is still not available.

With rotation of the cylinder(s), the problem is further complicated. As shown by the pressure data and confirmed by the flow visualization, the unsteady flow can be separating and reattaching over a large portion of the top surface. In the absence of any reliable procedure to account for wall confinement effects in the present situation, the results are purposely presented in uncorrected form.

**2.2a Base airfoil:** The pressure distribution data for the “base airfoil” (in absence of the modifications imposed by the leading-edge or upper-surface cylinder) are presented in figure 2. The leading edge was now formed by a snugly fitting plug (the nose fill-in section). Due to practical difficulties in locating pressure taps in the cusp

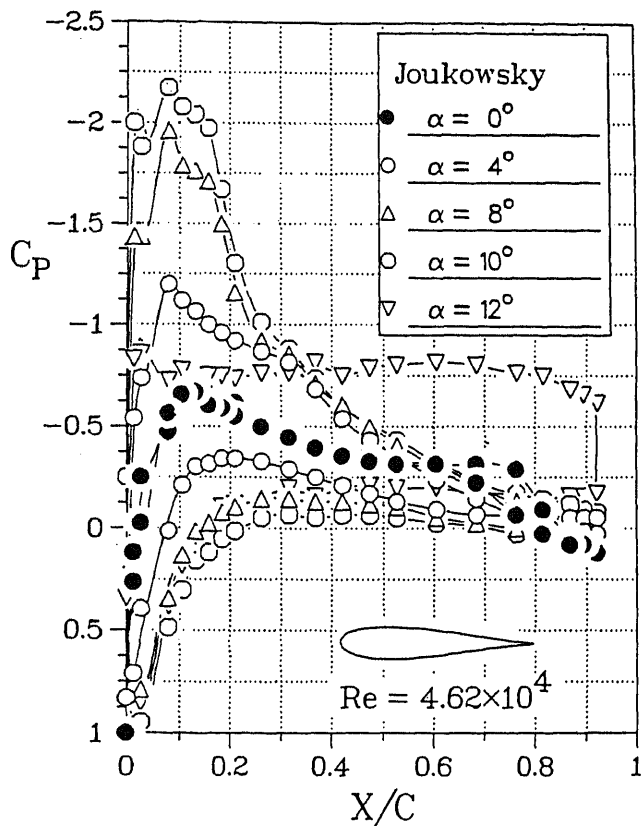


Figure 2. Experimentally obtained pressure distributions for the basic Joukowski model.

region, there is an apparent discontinuity in the pressure plots near the trailing edge. However, this region has little importance in the present discussion. It is apparent that the airfoil, in absence of any modifications to its nose geometry, stalls at an angle of attack of around  $10\text{--}12^\circ$ . These results serve as reference to assess the effect of rotating cylinders in different locations.

Note that the wall confinement effect at  $\alpha = 10^\circ$  is relatively small, as the blockage ratio is around 7%. More importantly, focus here is on the effect of the momentum injection due to the cylinder rotation with the airfoil at a given angle of attack. Results of the flow visualization study, presented later, emphasize this point.

**2.2b Leading edge cylinder:** Figure 3 summarizes the effects of modification of the airfoil with the leading-edge cylinder and the cylinder rotation. The base airfoil has a maximum lift coefficient of about 0.87 at an angle of attack of  $10^\circ$ . There is a penalty associated with the modified nose geometry as well as due to the gap, but even at the lowest rate of rotation of the cylinder ( $U_c/U = 1$ ) the lift and stall characteristics are significantly improved. The airfoil exhibits a desirable flattening of the lift curve at stall. The maximum lift coefficient measured with  $U_c/U = 4$  was around 2 at  $\alpha = 28^\circ$ , which is almost three times the lift coefficient of the base airfoil.

Typical pressure plots at a relatively larger angle of attack are presented in figure 4 to assist in more careful examination of the local flowfield. As the angle of attack of the airfoil is increased, the flow starts to separate from the upper surface close to the

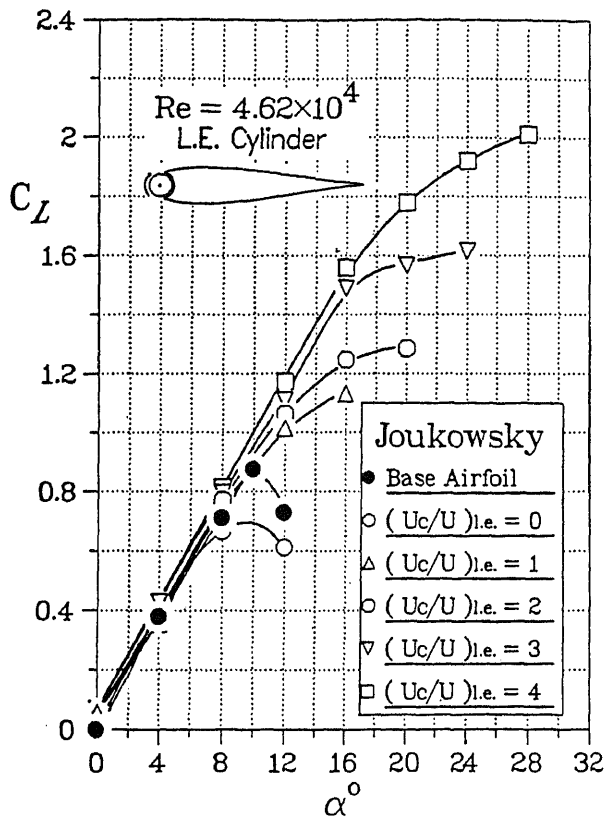


Figure 3. Effect of the leading-edge cylinder rotation on the lift and stall characteristics of the Joukowski model.

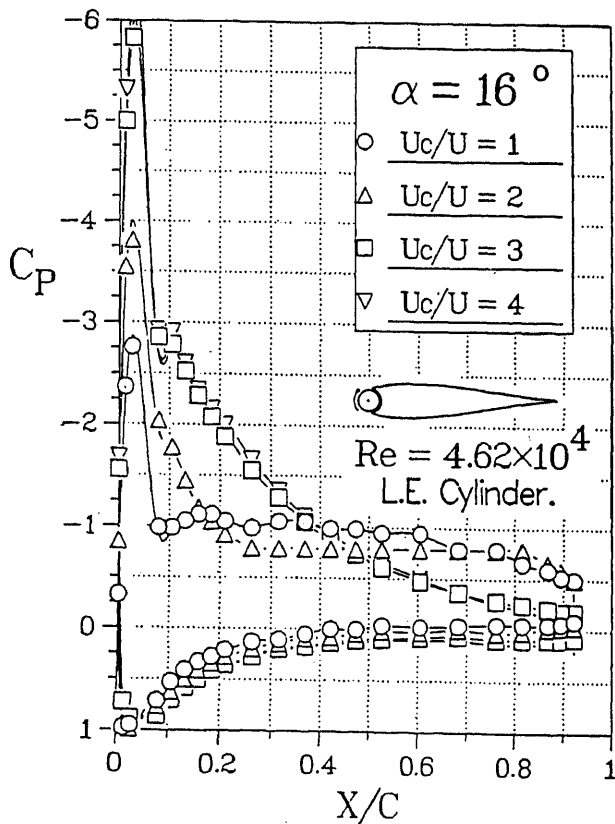
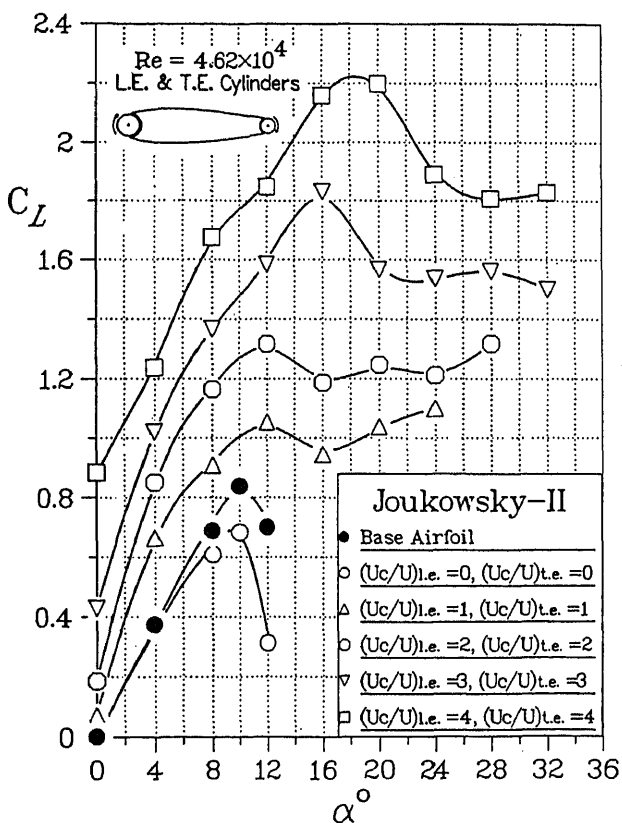


Figure 4. Effect of increasing the rate of cylinder rotation on pressure distribution around the model at a relatively larger angle of attack of  $\alpha = 16^\circ$

leading edge. At  $\alpha = 16^\circ$ , for example, the cylinder rotating at  $U_c/U = 1$  only keeps the flow attached at the leading-edge. However, as the rate of rotation is increased, the size of the separated region is reduced, and at the higher rates of rotation the flow is again completely attached. Note that the point of separation on the upper surface clearly moves downstream with an increase in rate of rotation. The flow separates at around  $X/C = 25\%$  with  $U_c/U = 2$ , near  $X/C = 80\%$  when  $U_c/U$  is increased to three, and at the trailing edge with the highest  $U_c/U$  used. The flow visualization study discussed later substantiated this general behaviour rather dramatically.

**2.2c Combined leading and trailing edge cylinders:** The use of a leading-edge cylinder extends the lift curve, thus substantially increasing the maximum lift coefficient and delaying stall. On the other hand, the trailing-edge cylinder rotation results in an improvement in the lift coefficient, at a given angle of attack, before stall. In order to combine these effects, the base configuration was modified to include both the leading and trailing-edge cylinders. This phase of the test-programme examined the effect of individual and combined cylinder rotations. However, it is the combined effect of both the cylinders that is of interest here. Results shown in figure 5 suggest some benefit due to rotation of the two cylinders together. Although the increase in the  $C_{L,max}$  is rather modest (from 2.0 to 2.5, around 30%) compared to the leading-edge cylinder case (sharp trailing edge, figure 3), the lift coefficient at a given  $\alpha$  is indeed increased significantly, as expected, due to the leftward shift of the plots. As



**Figure 5.** Variation of  $C_L$  vs  $\alpha$  for a modified Joukowski airfoil with leading and trailing-edge cylinders.



noted before, this is due to the added circulation by the trailing-edge cylinder. For example,  $C_L = 0.8$  at  $\alpha = 8^\circ$  and  $(Uc/U)$  i.e. = 3 (figure 3), whereas for the same angle of attack and  $(Uc/U)$  i.e. =  $(Uc/U)$  t.e. = 3 the corresponding  $C_L \approx 1.57$ , an increase of around 96%. Similarly,  $C_L = 1.55$  for  $\alpha = 16^\circ$  and  $(Uc/U)$  i.e. = 4. On the other hand, with both the cylinders rotating at  $Uc/U = 4$ , the lift coefficient is around 2.48, a further gain of about 60%. Note, the maximum lift coefficient attained with rotation of both the cylinders represents an increase of 195% with respect to the reference configuration ( $C_{L,max}$  of about 2.6 vs 0.88, figure 5).

**2.2d Forward and rear upper-surface cylinders:** The forward and rear upper-surface cylinders, located at 38 and 58% chord, respectively, were considered independently and with either operating in conjunction with the leading-edge cylinder. As can be expected, in the absence of rotation, their protrusion into the upper-surface flow had an adverse effect on the aerodynamic characteristics of the model. The flow separated at the location of the cylinder, resulting in lower lift and increased drag. On the other hand, with rotation, either of the upper-surface cylinders was successful in attaining a higher  $C_{L,max}$  and delaying the stall. In this respect, the forward upper-surface cylinder was particularly effective.

**2.2e Upper leading-edge cylinder:** Effectiveness of the combination of leading-edge and forward upper-surface cylinders suggested the possibility of replacing the two by a single rotating element. This avoids the practical complications associated with construction, installation and operation of two rotating cylinders.

The configuration, with a cylinder located at approximately 5% of the chord, was tested at cylinder speeds in the range of  $Uc/U$  up to 4. The results are presented in figure 6. Compared to the leading edge cylinder study (figure 3), where for  $Uc/U = 4$ ,

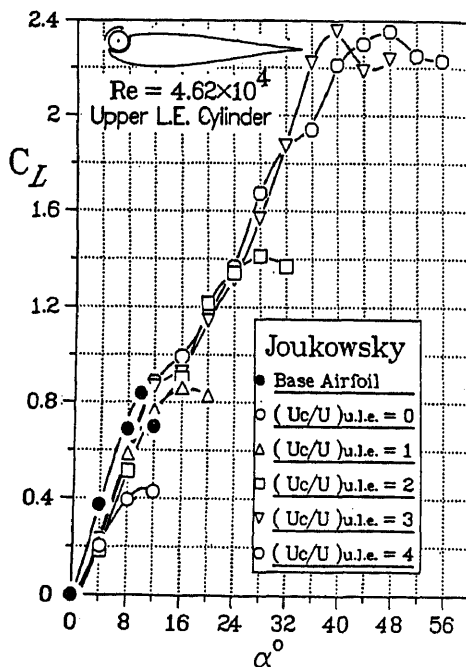


Figure 6. Lift and stall characteristics of the Joukowsky model as affected by the upper leading-edge cylinder rotation.

$C_{L,max} \approx 2$  and  $\alpha_{stall} \approx 28^\circ$ , now we have  $C_{L,max} = 2.35$  with  $\alpha_{stall} \approx 48^\circ$ . This clearly suggests that location of the cylinder near the leading edge can significantly affect the airfoil performance. Thus, there is room for a systematic study to arrive at an optimum location. Even compared to the results obtained using the leading-edge cylinder together with the forward upper-surface cylinder, performance of the present single cylinder configuration appears attractive. Although the  $C_{L,max}$  is slightly lower (down from 2.73 to 2.35), the stall is delayed from around  $40^\circ$ – $48^\circ$ . However, the main advantage would be the mechanical simplicity of working with one cylinder.

**2.2f Comparative performance:** With the vast amount of data obtained through a planned experimental program using the configurations presented earlier, it would now be useful to compare their distinctive features to help establish relative merits. Figure 7 attempts to achieve this objective. Results of the standard Joukowski airfoil (symmetrical, 15% thickness), with its  $C_{L,max} = 0.88$  and  $\alpha_{stall} = 10^\circ$ , serve as reference for all the cases presented.

The leading-edge cylinder is quite effective in extending the lift curve, without significantly changing its slope, thus substantially increasing the maximum lift coefficient ( $\approx 2$ ) and delaying the stall angle ( $28^\circ$ ). Further improvements in the maximum lift coefficient and stall angle are possible when the leading-edge cylinder is used in conjunction with an upper-surface cylinder. This configuration also results in lower drag due to large recovery of pressure near the trailing edge, at moderately high angles of attack. The  $C_{L,max}$  realized with the leading-edge and forward upper-surface cylinders, was about 2.73 ( $\alpha = 36^\circ$ ), approximately three times that of the base configuration.

A rotating cylinder on the upper side of the leading edge also proves to be very effective. Although the maximum coefficient of lift realized with its rotation is slightly

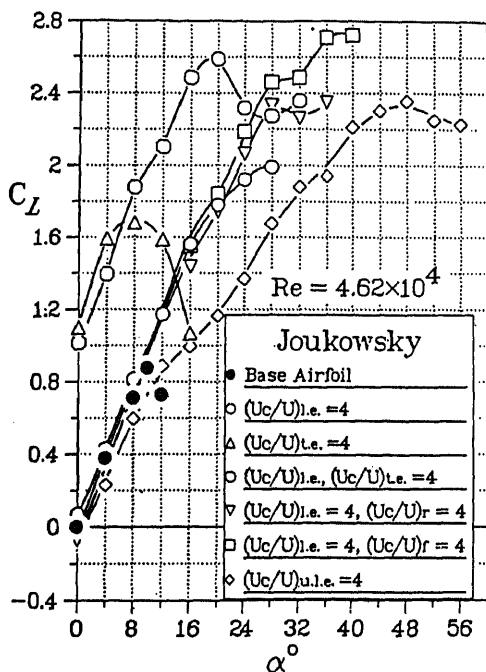


Figure 7. Plots to assess relative influence of different configurations studied on the lift and stall characteristics.

lower ( $\approx 2.35$ ), it does have a major advantage in terms of mechanical simplicity. Note that now the lift curve has a lower slope and is not an extension of the base airfoil lift curve. Hence, the lift at a given  $\alpha$  is relatively lower; however, the stall is delayed to around  $48^\circ$ .

On the other hand, to improve the lift over the range of low to medium angles of attack ( $\alpha \leq 20^\circ$ ), the trailing-edge cylinder proves much more effective, particularly in conjunction with the leading-edge cylinder. The suction over the airfoil upper surface as well as the compression on the lower surface are increased dramatically with the higher rates of rotation of this cylinder, resulting in a substantial increase in lift ( $\approx 195\%$ ).

Thus, depending on the intended objective in terms of desired  $C_{L,\max}$  and stall angle, one can select an appropriate configuration to initiate a preliminary design.

### 2.3 Flow visualization

The flow visualization study was carried out in a closed-circuit water channel facility. The model was constructed from Plexiglas and fitted with a leading-edge cylinder, driven by a compressed-air motor. A suspension of fine polyvinyl chloride powder was used in conjunction with slit lighting to visualize streaklines. Both angle of attack and cylinder speeds were systematically changed and still photographs as well as a video movie were taken. The study showed, rather dramatically, the effectiveness of this form of boundary-layer control (figure 8). With the model at  $\alpha = 20^\circ$ , and in absence of the cylinder rotation, a well-defined early separation resulting in a wide wake is quite apparent, with large-scale vortices sweeping away downstream. However, with the cylinder rotating at  $U_c/U = 4$ , an essentially attached flow is established over most of the upper surface of the airfoil.

At relatively lower rates of cylinder rotation, the flow character was found to be similar to that observed at  $U_c/U = 1$ , with the separation and reattachment regions progressively shifting downstream as the rotation rate increased. This is apparent through a progressive increase in  $U_c/U$  from 0 to 4. In fact, the flow pattern was found to be quite unsteady with the vortex layer separating and forming a bubble on reattachment, the whole structure drifting downstream, diffusing, and regrouping at different scales of vortices. Ultimately the flow sheds large as well as small vortices. This unsteady character of the separating shear layer and the wake is clearly evident in the video. Thus the flow character indicated by the experimentally obtained time-average pressure plots appears to be a fair description of the process. Furthermore, this also suggests that analytical or numerical modelling of such highly complex and transient flow would pose a challenging problem.

### 2.4 Numerical approaches

The complex problem of multi-element airfoil with momentum injection was studied using two distinctly different numerical approaches:

- (a) surface singularity distribution with boundary-layer correction;
- (b) finite element integration of the Navier–Stokes equations.

**2.4a Surface singularity approach:** This numerical procedure is based on the surface singularity method described in detail by Mokhtarian (1988) in his doctoral dissertation. It accounts for the wall confinement and involves replacement of the

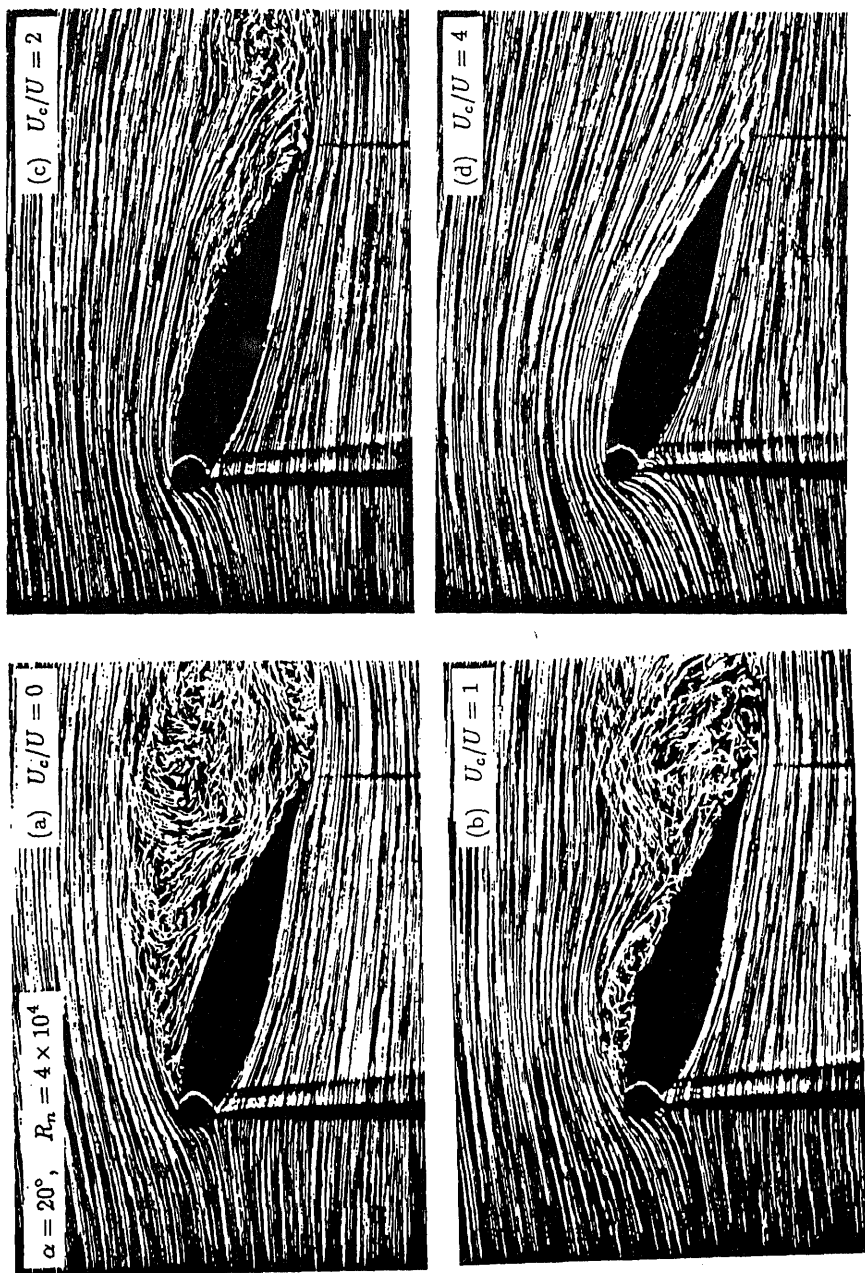


Figure 8. Typical flow visualization photographs showing the remarkable effectiveness of the moving surface boundary-layer control method: (a) highly separated flow, at a large angle of attack, in absence of the cylinder rotation; (b) appearance of a separation bubble,  $U_c/U = 1$ ; (c) downward shift of the separation point with a distinct reduction in the wake width,  $U_c/U = 2$ ; (d) essentially attached flow,  $U_c/U = 4$ .

aerofoil and wind tunnel walls with vorticity distribution  $\gamma$  in conjunction with appropriate constraint relations. Inclusion of a source within the contour of the airfoil models the wake when there is flow separation from the surface. A finite difference boundary-layer scheme is used to introduce viscous corrections. The scheme employs potential flow pressure distribution results to calculate the boundary-layer characteristics at the top and bottom surfaces starting from the stagnation point until the point of separation.

The procedure uses the displacement thickness to construct an equivalent airfoil and then iterates between the potential flow and boundary-layer scheme to converge to the final pressure distribution. Thus the objective is to match the outer potential flow solution with the inner boundary-layer prediction. The thin shear layer approximations of the Navier-Stokes equations for steady, two-dimensional, incompressible flow are used. The finite difference method employed for viscous correction is due to Keller & Cebeci (1972). The eddy viscosity term is expressed as suggested by Cebeci & Smith (1974) who treat the turbulent boundary-layer as a composite layer consisting of inner and outer regions with separate expressions for eddy viscosity in each region. The details of the formulation and the finite difference procedure followed are those given by Cebeci & Bradshaw (1977).

Typical results for the Joukowski airfoil with upper leading edge cylinder are presented in figure 9. Wind tunnel test results are also included to facilitate comparison. Considering the complex character of the flow, the correlation is indeed excellent and the results can be used with confidence.

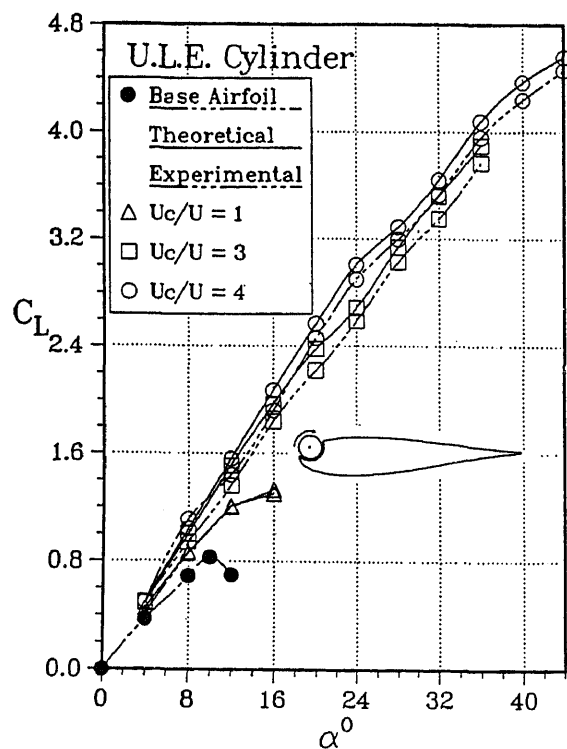
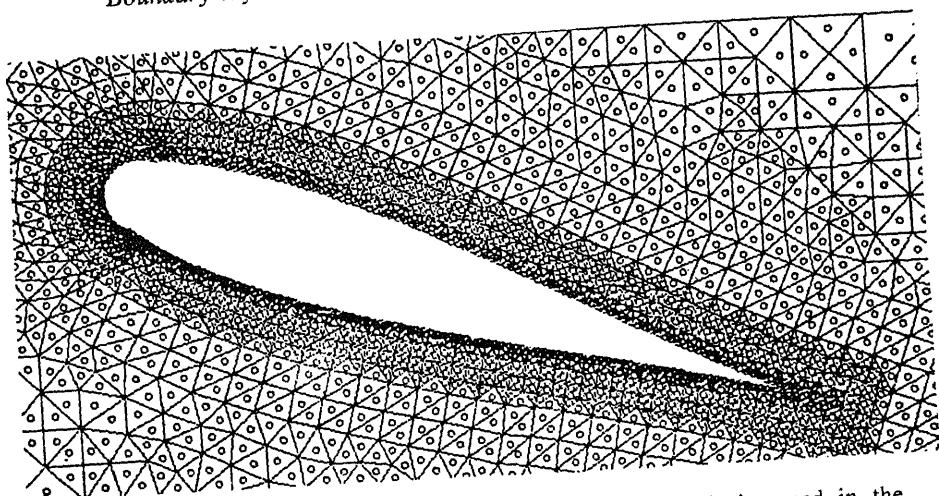
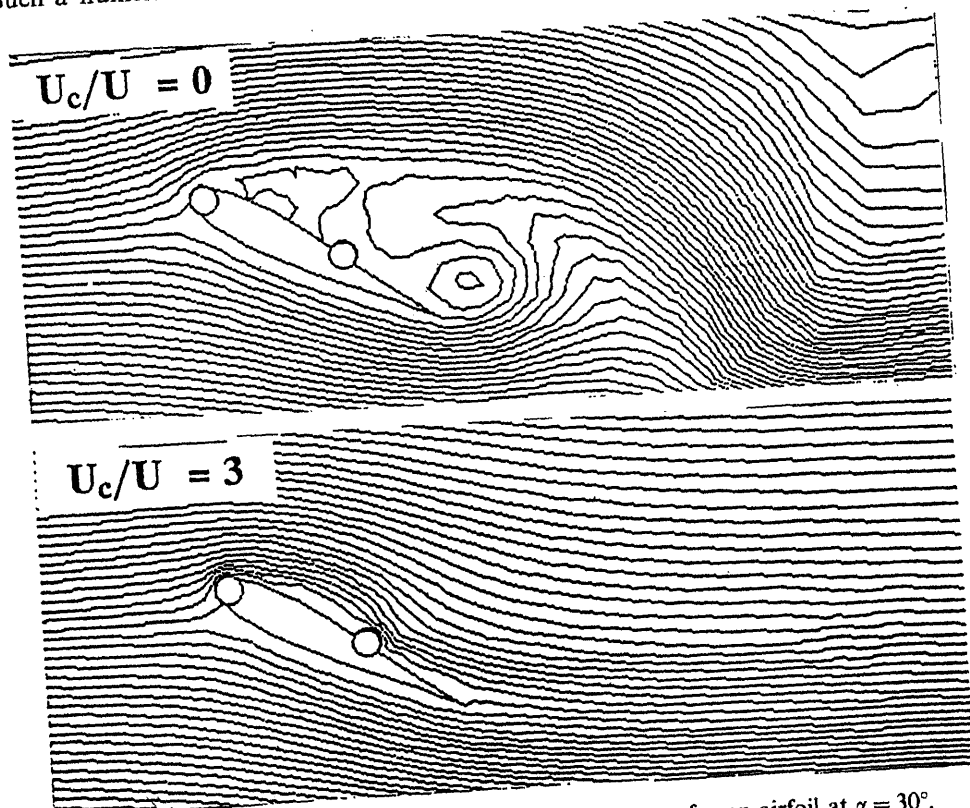


Figure 9. Variation of lift with angle of attack in presence of the MSBC as predicted by numerical and experimental procedures. Note that in spite of the complex character of the flow the correlation is excellent.



**Figure 10.** Triangular elements with spatially varying grid size used in the numerical simulation.

**2.4b Finite element Navier-Stokes solution:** Here the stream function-vorticity form of the Navier-Stokes equations are used in conjunction with the variable grid-size ( $\approx 3000$  nodes, figure 10) finite element analysis (Modi & Yokomizo 1992, pp. 270-4). Such a numerical solution of the two- and three-element airfoils with momentum



**Figure 11.** Typical numerically obtained flow patterns, for an airfoil at  $\alpha = 30^\circ$ , showing effectiveness of the MSBC.

injection has not been reported in literature. The parametric analysis involving a systematic variation of the speed ratio, angle of attack and the Reynolds number gave detailed information about the pressure loading, separation condition and the time dependent wake (figure 11). It also showed, rather spectacularly, the effectiveness of the MSBC.

### 3. Flat plate and rectangular prisms

#### 3.1 Model and wind tunnel test-programme

Two-dimensional flat plate and rectangular prisms were tested in a  $45 \times 45$  cm cross-section wind tunnel with a maximum speed of 50 m/s. The large converging nozzle at the entrance of the tunnel (contraction ratio = 10:1) made the flow in the test-section uniform with a level of turbulence less than 0.5%. The tunnel speed was adjusted by a variac transformer and measured using a pitot static tube connected to an inclined alcohol manometer.

A flat plate model,  $9 \times 40.5$  cm, and two rectangular prism models with a span ( $B$ ) of 40.5 cm and depth ( $L$ ) to width ( $H$ ) ratio  $L/H = 0.3, 1, 2, 4$  were constructed from Plexiglas. The models were equipped with two moving surface boundary-layer control elements (rotating cylinders) as shown in figure 12. The cylinders were driven by variac controlled AC motors through flexible belt drives. The motor speed was monitored using a strobe light. In the present test-programme the ratio  $U_c/U$  was varied from 0–3. This corresponded to a maximum cylinder speed of around 11,000 rpm at a free stream speed of 5 m/s. To ensure two-dimensionality of the flow the models were fitted with end plates. In general, the tunnel speed was kept constant at 5 m/s, which corresponds to a Reynolds number of  $3 \times 10^4$  based on the free stream velocity and the model width ( $H$ ). The lift and drag forces as well as pressure data were recorded over a range of the angle of attack at  $5^\circ$  increments. The force could be measured with an accuracy of 0.5 gm/mV. Details of the test-arrangement and results are discussed at length in earlier publications (Modi *et al* 1989, 1990a, 1991b,c).

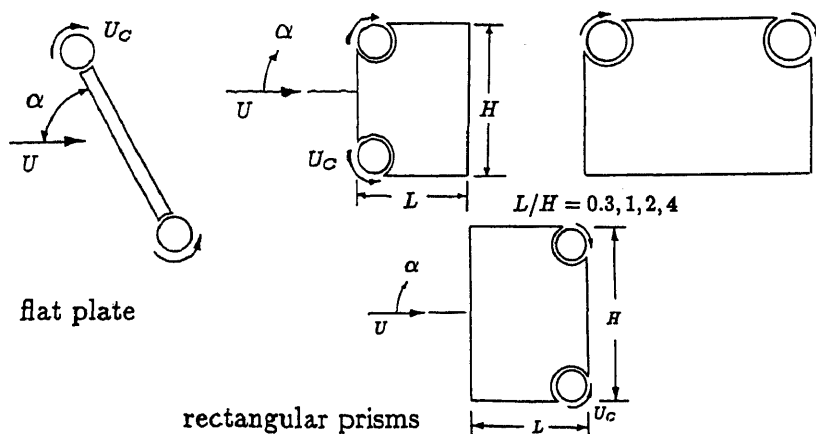


Figure 12. Schematic diagrams of the two-dimensional bluff bodies used during the wind tunnel tests and flow visualization.

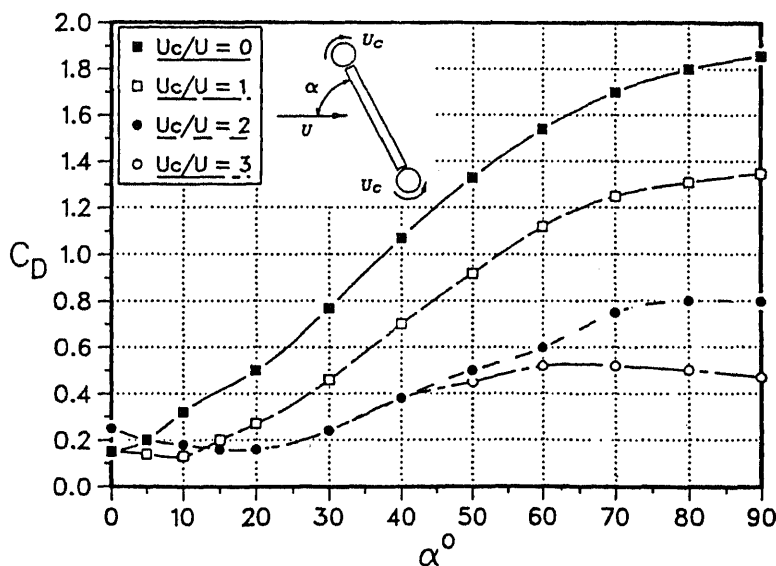
### 3.2 Results and discussion

**3.2a Flat plate:** Tests with the flat plate were carried out with either of the cylinders rotating independently; or with the two cylinders rotating together, but in the opposite sense, for effective momentum injection to assist in the boundary-layer control. Both the lift as well as the drag results showed remarkable improvement (Modi *et al* 1990b,c).

Of course, the maximum reduction in wake and hence the corresponding decrease in the drag coefficient can be expected when both the cylinders are rotating as shown in figure 13. For  $\alpha = 90^\circ$ , a decrease in the drag coefficient from 1.85 at  $U_c/U = 0$  to 0.47 at  $U_c/U = 3$  represents a reduction of around 75%. The flow visualization photographs also showed a remarkable reduction in the wake width thus qualitatively substantiating the trend suggested by the wind tunnel test results.

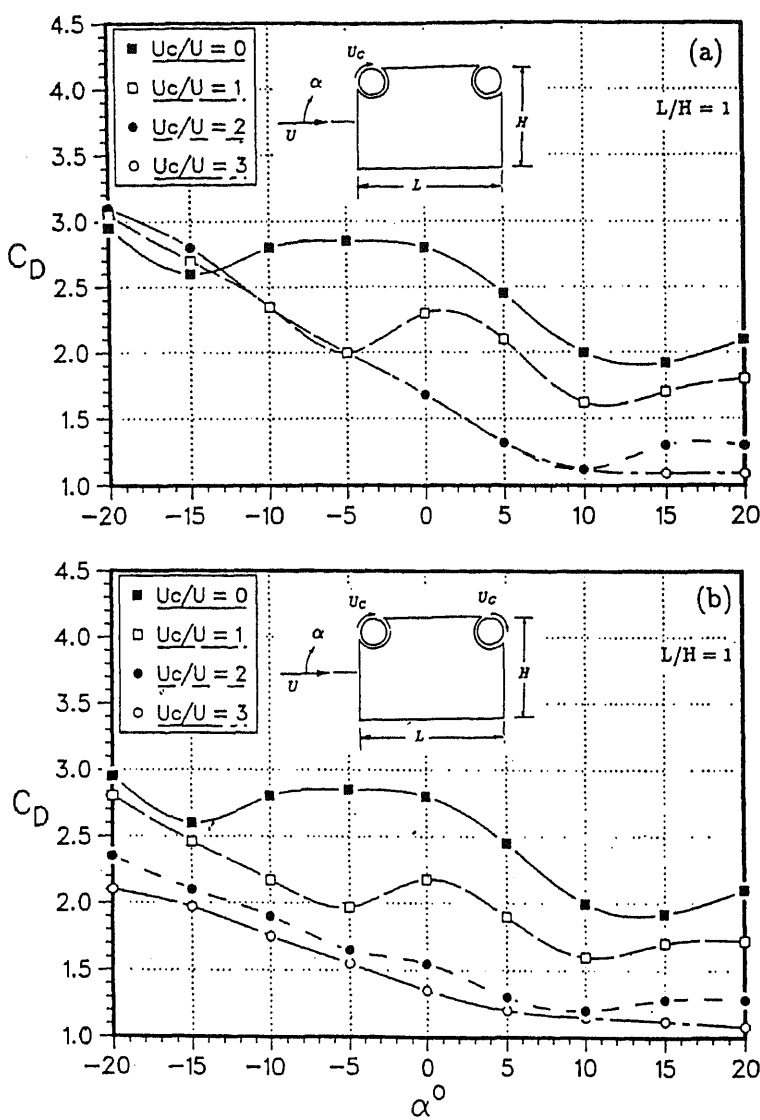
**3.2b Rectangular prisms:** Rectangular prisms with rotating cylinders at two adjacent corners provide three basic configurations for study: the side with cylinders facing the flow, forming the top face, or representing the rear face. Various intermediate configurations can be obtained by systematically changing the angle of attack. With four values of  $L/H$ , to help assess the effect of boundary-layer reattachment and reseparation further downstream, and four values of  $U_c/U$ , the amount of information obtained is rather extensive. Only some typical results in discerning trends are presented here.

Figure 14 shows a sample of the representative results for the experimental phase where the rotating elements are on the top surface, i.e. parallel to the free-stream for  $\alpha = 0$ . Cases corresponding to single- and two-cylinder rotation for the square prism model are considered. At the outset it is apparent that rotation of the second cylinder has very little effect on the flow field, and hence on  $C_D$ , for  $\alpha > 5^\circ$ , as now the trailing



**Figure 13.** Plots showing significant reduction in drag of a two-dimensional flat plate with the moving surface boundary-layer control applied at both the leading and trailing edges. Note that at  $\alpha = 90^\circ$ , the reduction in  $C_D$  is around 75%.





**Figure 14.** Variation of the drag coefficient with the angle of attack for a two-dimensional square prism when the boundary-layer control is applied at the top surface: (a) a rotating cylinder at the leading edge; (b) rotating cylinders at leading and trailing edges.

edge cylinder lies in the wake. However, for smaller and negative  $\alpha$  it is quite successful in further reducing  $C_D$ : from 1.7 at  $\alpha = 0^\circ$  and  $U_c/U = 3$  for the upstream cylinder rotation to 1.3 when both the cylinders are rotating. A reduction in the drag coefficient by 54% with both the cylinders rotating is indeed quite impressive.

The influence of rotating cylinders located on the rear vertical face of the square and rectangular prisms was also investigated. In this case, the boundary-layer separates at the top and bottom leading edges and the rotating cylinders are

submerged in the wake thus reducing their effectiveness. Now the reduction in  $C_D$  at  $\alpha = 0^\circ$  and  $U_c/U = 3$  was found to be only 13% for the square prism and virtually zero for the rectangular prism compared to 53 and 40%, respectively, for the case with cylinders at the front face.

## 4. Tractor-trailer truck configuration

### 4.1 Background

A comprehensive literature review of the road vehicle aerodynamics suggests that although aerodynamically contoured car design has become standard practice lately, trucks and buses have changed little during the past 30 years (Sovaran *et al* 1978; Kramer & Gerhardt 1980). Most of the modifications have been limited to rounded edges with provision for vanes, skirts and flow deflectors. The benefit due to some of the "add-on" devices is still a matter of controversy and, at best, marginal under conditions other than the specific ones used in their designs. Bearman (1980) has presented an excellent review on the subject (with 54 references cited). The thesis by Wacker (1985) also discusses the limited influence of "add-on" devices with a possibility of increasing the drag under non-optimal conditions. On the other hand, it was found that judicious choice of ground clearance, gap-size between the tractor and trailer, and back inclination can reduce the drag coefficient by a significant amount.

A word concerning numerical analysis of the complex aerodynamics associated with road vehicles would be appropriate. A reliable and cost-effective methodology, if available, can assist in design with reduced dependence on time-consuming and expensive wind tunnel tests. With the advent of supercomputers, parallel processing and neural network concepts, considerable progress has been made in that direction. However, modelling of three-dimensional boundary layers around a complex geometry at supercritical Reynolds numbers, with separation, reattachment and reseparation of unsteady turbulent flows still represents a challenging problem (Kataoka *et al* 1991).

### 4.2 Model and test-procedure

A 1/12 scale tractor-trailer truck model was constructed out of Plexiglas. The model has a trailer with width  $B = 22.7$  cm, height  $H = 26.2$  cm, length  $L = 128.4$  cm, and a hydraulic diameter of 0.311 m. A typical truck model was supported by four steel guy wires which were suspended from the ceiling and carried turnbuckles to help level the model. As the length of the wire ( $\approx 145$  cm) is much larger than the maximum horizontal displacement of the truck model ( $\leq 5$  cm), the drag induced displacement was essentially linear in the downstream direction.

Variation in the drag, due to the boundary-layer control devices being relatively small, required development of a sensitive transducer for its measurement. The drag-induced downstream motion of the model was transmitted by an inelastic string to a cantilever beam with a pair of strain gauges near its root. The gauges formed a part of the Wheatstone Bridge (of the *Bridge Amplifier Meter, BAM*) and the amplified filtered output was recorded using a DISA voltmeter. The sensitivity of the drag measurements was around 0.4 g/mV.

### 4.3 Results and discussion

Tests with a scale model of the truck were carried out in the boundary-layer tunnel with negligible blockage effect (blockage ratio = 1.2%). The trailer was provided with rotating cylinders at its top leading edge and downstream locations. The  $L/H$  ratio for the trailer was approximately 3.75 which suggested that rotation of the rear cylinder has virtually no effect on the drag reduction. The wind tunnel tests substantiated this observation. Considering the fact that:

- (i) around 70% of goods in North America are transported by trucks;
- (ii) depending upon the speed, approximately 40–70% of the power is expended in overcoming the aerodynamic drag;
- (iii) on an average, a truck travels around 150,000 km/year;

even 1% reduction in the drag coefficient can translate into substantial savings in fuel costs.

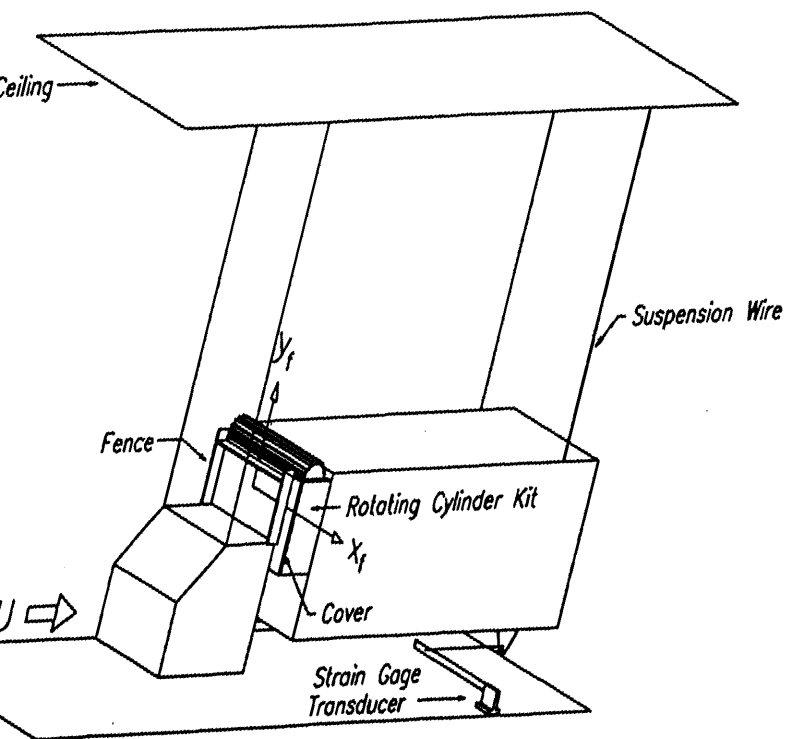
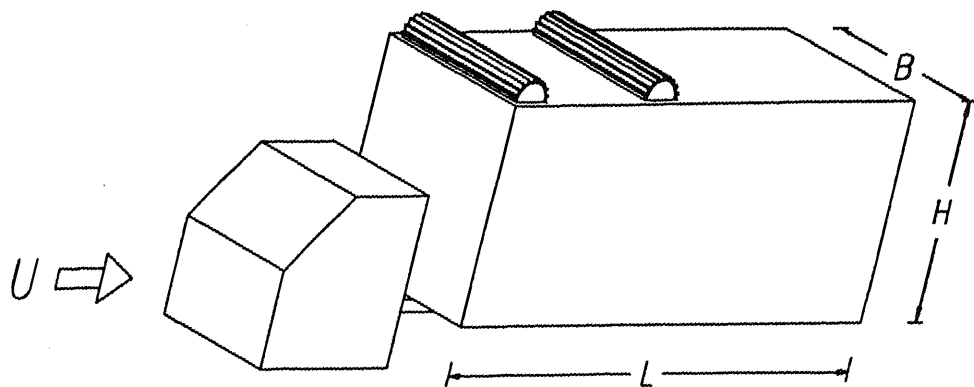
With the positive influence of the cylinder roughness on the momentum injection process and associated reduction in drag, it seemed logical to introduce the momentum more directly. This was achieved in several ways:

- (i) provide increased cylinder surface roughness through roughness squares, helical grooves or splines running parallel to the cylinder axis;
- (ii) keep one cylinder at the top leading edge of the trailer (referred to as the front cylinder) and locate the second cylinder (rear cylinder) at an optimum distance downstream. The objective is to inject additional momentum in the boundary-layer to compensate for dissipation of the momentum introduced by the front cylinder and thus counter the emergence of adverse pressure gradient;
- (iii) raise the cylinders so as to immerse them in the boundary layer and assess the effect of cylinder orientation.

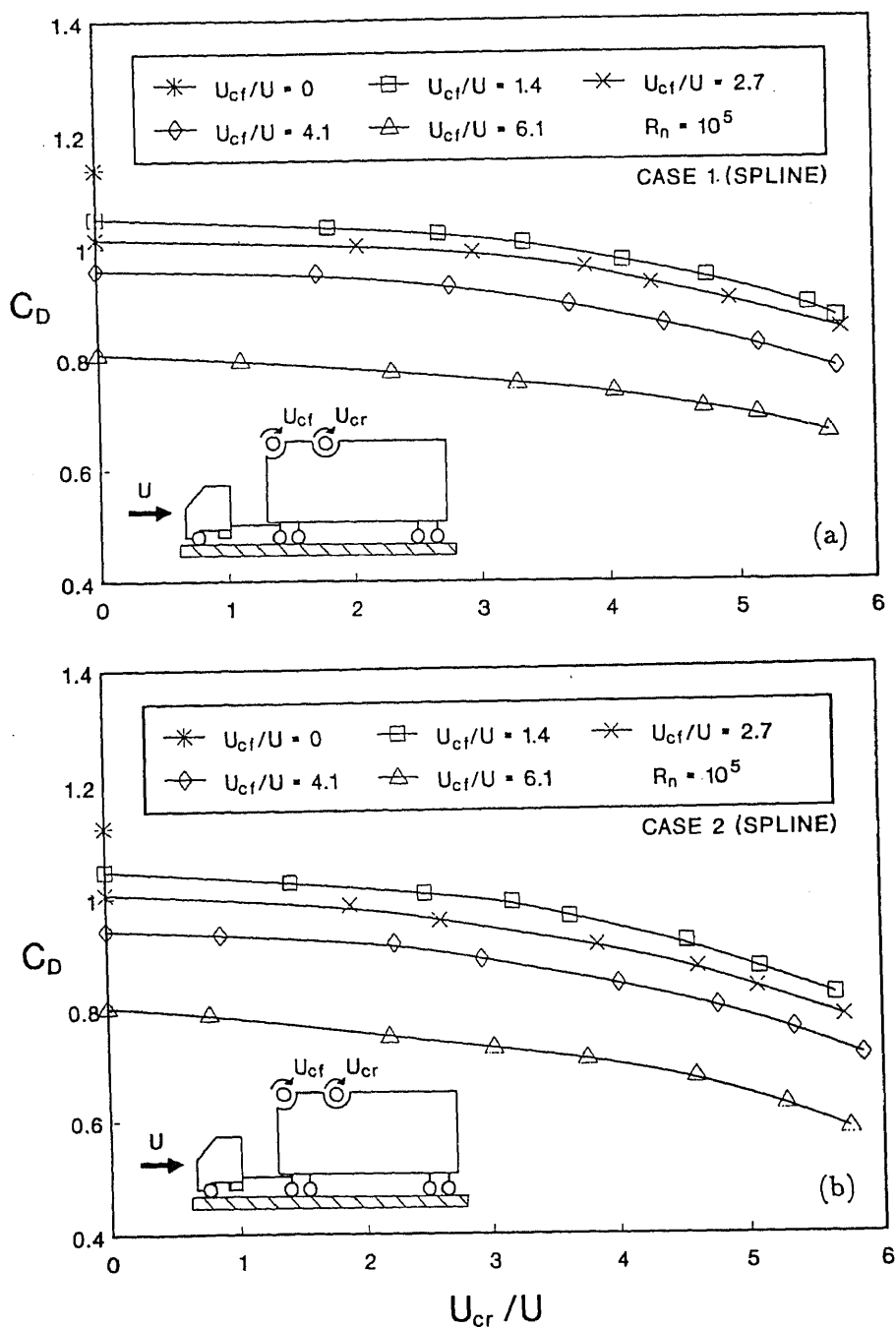
Extensive wind tunnel tests with different combinations of speed ratio ( $U_c/U$ ), cylinder location and surface roughness showed the helical groove and spline geometry, with one cylinder located at the leading edge and the other 25.4 cm downstream, to

**Table 1.** Wind tunnel tests conducted with different speed ratios and orientation of the twin helical groove and spline cylinders. The front cylinder is located at the top leading edge of the trailer. The second cylinder is located 25.4 cm (10 in) downstream.

Case	Cylinder location	
	Front raised (mm)	Rear raised (mm)



Schematic diagram of the 3-D tractor-trailer truck model and its instrumentation in the UBC's boundary-layer wind tunnel.



**Figure 16.** Variation of the drag coefficient  $C_D$  with the speed ratio for the spline twin-cylinder configuration: (a) Case 1: both cylinders flush; (b) Case 2: front cylinder flush, rear cylinder raised 6.35 mm.

be quite effective. The effect of raising the cylinder above the trailer surface was also found to be significant. The cylinder orientations studied with helical and spline roughnesses are indicated in table 1. Figure 15 schematically shows the model and the test arrangement. The extensive amount of information obtained has been reported elsewhere (Modi *et al* 1991c, pp. 465–82; Ying 1991). Only some typical results are presented here to indicate its potential.

Figure 16a shows the effect of the spline cylinder rotation for case 1, i.e. when both the cylinders are flush with the top face of the trailer.  $U_{cf}$  and  $U_{cr}$  refer to front and rear cylinder surface speeds, respectively. At the outset it is apparent that the front cylinder rotation ( $U_{cr} = 0$ ) reduces the drag coefficient rather significantly, from 1.14 at  $U_{cf} = 0$  to 0.96 at  $U_{cf} = 4.1$ , a drop of around 15.8%. Rotation of the rear cylinder improves the situation further and for both the cylinders with a speed ratio  $U_c/U = 4.1$ , the reduction in  $C_D$  reaches 22.8%!

The effect of raising the rear cylinder is shown in figure 16b. Note that even in absence of the momentum injection ( $U_{cf} = U_{cr} = 0$ ), the reference drag coefficient is slightly reduced ( $C_D = 1.19$ ). This may be attributed to the combined effect of an increase in the projected area on which the drag coefficient is based and the large wake width caused by the rear cylinder. Rotation of the front cylinder does not seem to improve the situation significantly (compared to case 1), as for  $U_{cf}/U = 4.1$ , the reduction in drag is 16.8%. With both the cylinders rotating at a speed ratio of 4.1, the decrease in drag coefficient amounts to 24.8%.

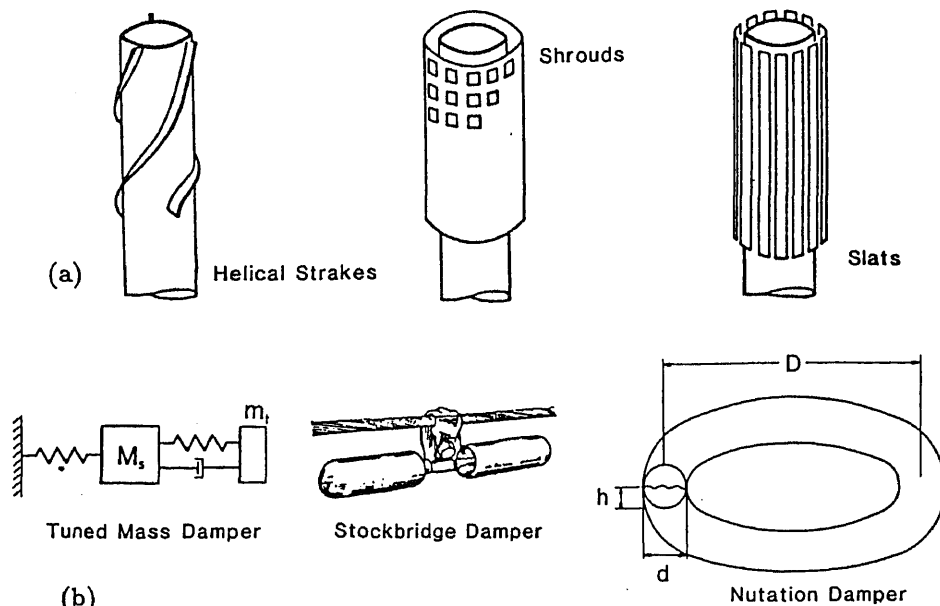
Essentially the same trend continued to persist when the front cylinder was also raised (case 7). The drag coefficient in absence of the cylinder rotation dropped further to 1.12 as explained before. With  $U_{cf}/U = U_{cr}/U = 4.1$ , the reduction in drag reached almost 26%. Thus the splined geometry of the rotating elements with raised positions, appears quite promising in reducing the pressure drag of the tractor-trailer truck configuration through MSBC.

## 5. Control of wind-induced instabilities

With the success of the moving surface boundary-layer control, in increasing lift and reducing drag of both slender bodies at high angles of attack and bluff geometries, attention was directed towards control of wind induced instabilities.

The response of aerodynamically bluff bodies when exposed to fluid streams has been a subject of considerable study for quite some time. The prevention of aeroelastic vibrations of smokestacks, transmission lines, suspension bridges, tall buildings etc. is of particular interest to engineers. Ever since the pioneering contribution by Strouhal, who correlated periodicity of the vortex shedding with the diameter of a circular cylinder and the velocity of the fluid stream, there has been a continuous flow of important contributions resulting in a vast body of literature. This has been reviewed rather adequately by Cermak (1975), Modi & Slater (1991), Welt (1988) and others. In general, the oscillations may be induced by vortex resonance or geometric-fluid dynamic instability called galloping.

Several passive devices such as helical strakes, shrouds, slats, tuned mass and nutation dampers etc. have been proposed over the years (figure 17) and have exhibited varying degrees of success in minimizing the effects of vortex induced and galloping types of instabilities (Zdravkovich 1980). In general, vibration suppressing devices tend to change the aerodynamic characteristics of the structure in such a way as to



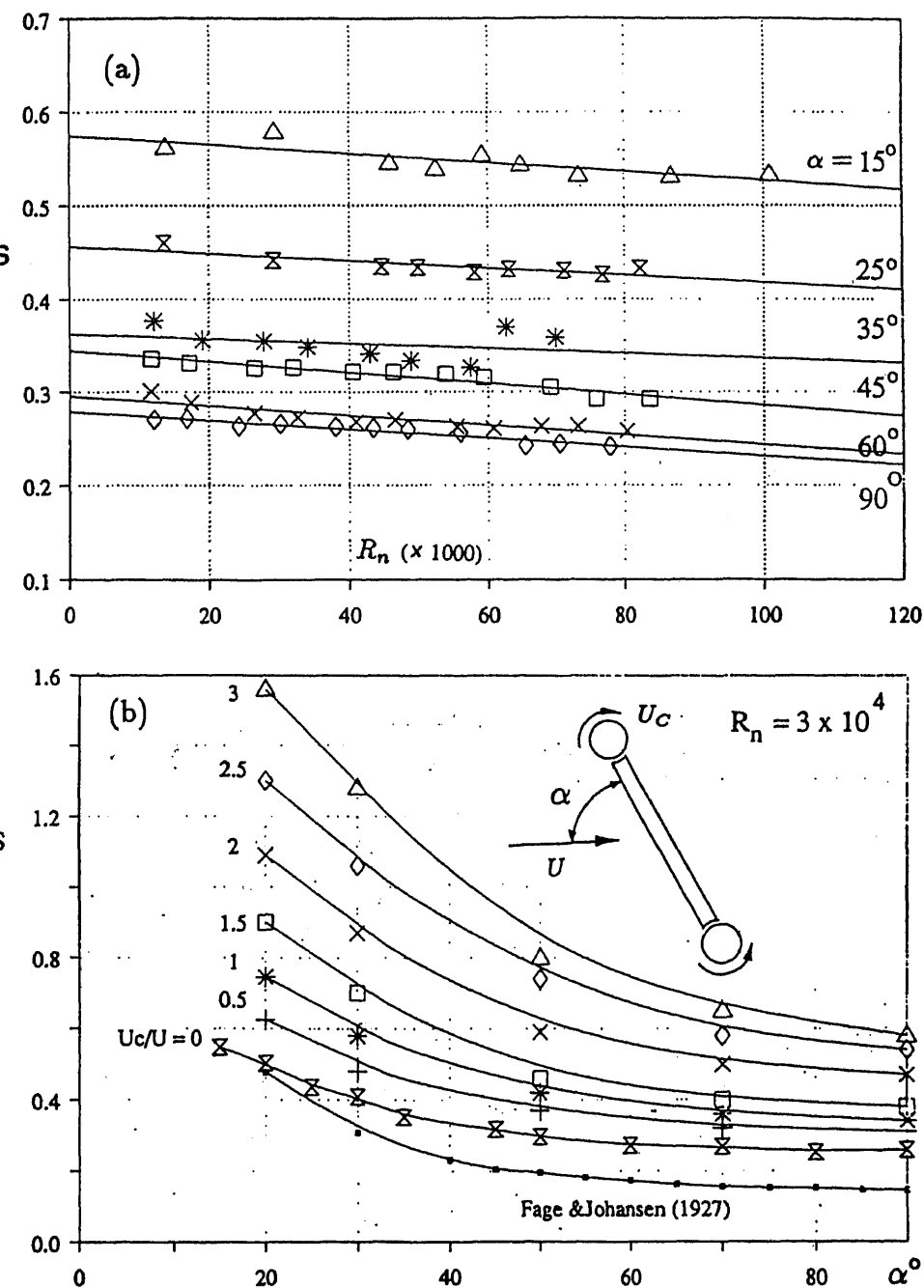
**Figure 17.** Passive devices used to control wind induced instabilities: (a) strakes, slats, and shrouds modify system aerodynamics; (b) dampers provide energy dissipation mechanism.

interfere with and weaken the existing force; while the dampers provide a mechanism for dissipating energy. It is of interest to note that all the above mentioned procedures are passive in character. Semi-active devices such as a rotating element for the boundary-layer control and, through it, damping of the instability has received virtually no attention. Such applications of the MSBC were explored only recently (Kubo *et al* 1991; Modi *et al* 1991d).

Variation of the Strouhal number ( $S$ ) with Reynolds number ( $Re$ ) for the flat plate with rounded tips due to the presence of cylinders but in the absence of their rotation is shown in figure 18a. As expected, at a given Reynolds number,  $S$  diminishes as the angle of attack increases. Effect of the cylinder rotation on the Strouhal number as a function of  $\alpha$  for a fixed  $Re$  of  $3 \times 10^4$  is presented in figure 18b. The classical results of Fage & Johansen (1927) are also included for comparison. A remarkable increase in the Strouhal number with cylinder rotation at a given  $\alpha$  is apparent. It suggests an increase in the shedding frequency corresponding to narrowing of the wake. Obviously, this will delay the onset of vortex resonance.

Kubo *et al* (1991) applied the concept to a two-dimensional square prism provided with twin rotating elements at the corners of the front face. The tests were carried out in a wind tunnel with the spring supported model free to undergo plunging oscillations. The single cylinder rotation affected the coherence of the vortex shedding and hence suppressed resonant instability. On the other hand, cylinder rotation successfully modified the loading to arrest the galloping. Excellent flow visualization pictures supported the wind tunnel test observations.

After an uncertain beginning and interrupted advances over nine decades, it is



**Figure 18.** Strouhal number ( $S$ ) associated with a flat plate having rounded edges due to tip cylinders: (a) variation with Reynolds number and angle of attack in the absence of cylinder rotation; (b) effect of momentum injection.



apparent that the field now presents an exciting opportunity for contributions. As the Gita puts it:

*"Knowledge is merely a small island surrounded by a vast ocean of ignorance"*

No matter how far we advance, we will always be on the shores of that uncharted ocean. But then, a journey fulfills itself at every step.

## 6. Concluding remarks

Based on a rather fundamental study of moving surface boundary-layer control with two-dimensional airfoils, plates and prisms, as well as its application to a scale model of a typical tractor-trailer truck configuration, the following conclusions can be made:

- (i) Moving surface boundary-layer control (MSBC) can significantly increase lift, decrease drag and delay stall of an aircraft. Its application to the next generation of high performance airplanes is indeed quite exciting. NASA and the US Air Force are actively looking into this aspect;
- (ii) the concept also appears to be quite promising in reducing the drag of bluff bodies. For the flat plate at  $\alpha = 90^\circ$ , it reduced the drag coefficient by 75%. The maximum reduction for a square prism varied from 54% ( $L/H = 1$ ) to 40% ( $L/H = 2$ );
- (iii) effectiveness of the momentum injecting device diminishes when located in the wake;
- (iv) surface roughness of the rotating cylinder tends to improve the boundary-layer control;
- (v) the MSBC concept also proved effective in reducing the drag of a truck configuration by 26%;
- (vi) the concept is essentially semi-passive in character requiring a negligible amount of power for its implementation;
- (vii) numerical approach to the problem using the surface singularity procedure, as well as finite element and finite difference application to integration of the Navier-Stokes equations, presents considerable scope for contribution.
- (viii) flow visualization study confirms the effectiveness of the MSBC quite dramatically.
- (ix) application of the concept in arresting wind-induced vortex resonance and galloping instabilities appears quite promising.

The models were fabricated in the Mechanical Engineering Workshop. The assistance of M/s E Abell, P Hurren and D Camp in the design and construction of the models is gratefully acknowledged. The investigation was supported by the Natural Sciences and Engineering Research Council of Canada, Grant No. A-2181.

## References

- Bearman PW 1980 Review of bluff body flows applicable to vehicle aerodynamics. *Trans. ASME, J. Fluids Eng.* 102: 265-274
- Betz A 1961 History of boundary layer control in Germany. *Boundary layer and flow control* (ed.) G V Lachmann (London: Pergamon) vol. 1

- Cebeci T, Bradshaw P 1977 *Momentum transfer in boundary layer* (Washington: Hemisphere-McGraw Hill)
- Cebeci T, Smith A M O 1974 *Analysis of turbulent boundary layers* (New York: Academic Press)
- Cermak J E 1975 Application of fluid mechanics to wind engineering: Freeman Scholar Lecture. *Trans. ASME, J. Fluids Eng.* 97: 9-38
- Chang P K 1970 *Separation of flow* (New York: Pergamon)
- Fage A, Johansen F C 1927 On the flow of air behind and inclined flat plate of infinite span. *Proc. R. Soc. London A* 116: 170-197
- Favre A 1938 *Contribution a L'etude Experimentale des Mouvement Hydrodynamiques a Deux Dimensions*, doctoral thesis presented to the University of Paris
- Flettner A 1925 The Flettner rotor ship. *Engineering* 19: 117-120
- Goldstein S 1938 *Modern developments in fluid mechanics* (Oxford: University Press) vols. 1 & 2
- Iverson J D 1972 Correlation of magnus force data for slender spinning cylinders. *AIAA 2nd Atmospheric Flight Mechanics Conference*, Palo Alto, California, Paper No. 72-966
- Kataoka T, China H, Nakagawa K, Yanagimoto K, Yoshida M 1991 Numerical simulation of road vehicle aerodynamics and effect of aerodynamic devices. *SAE International Congress and Exposition*, Detroit, Paper No. 91-0597
- Keller H B, Cebeci T 1972 Accurate numerical methods for boundary-layer flows, Part 2. Two-dimensional turbulent flow. *AIAA J.* 10: 1193
- Kramer C, Gerhardt H J 1980 Road vehicle aerodynamics. *Proceedings of the 4th Colloquium on Industrial Aerodynamics* (Aachen, Germany: Fachhochschule)
- Kubo Y, Modi V J, Yasuda H, Kato K 1991 On the suppression of aerodynamic instabilities through the moving surface boundary-layer control. *Proceedings of the 8th International Conference on Wind Engineering* (ed.) A G Davenport (London, Ont: University of Western Ontario)
- Lachmann G V 1961 *Boundary layer and flow control* (London: Pergamon) vols. 1 & 2
- Magnus G 1853 Ueber die Verdichtung der Gase an der Oberflache Glatter Korper. *Poggendorfs Ann. Phy. Chem.* 88: 604-610
- Modi V J 1991 Moving surface boundary-layer control: Experiments, analysis and applications. *The G I Taylor Memorial Lecture 36th ISTAM Congress* (Bombay: Indian Inst. Technol India)
- Modi V J, Fernando M S U K, Yokomizo T 1989 Moving surface boundary layer control as applied to two- and three-dimensional bodies. *Proceedings of the 8th Colloquium on Industrial Aerodynamics-Industrial Flows* (eds.) C Kramer, H Gerhardt (Aachen, Germany: Fachhochschule) pp. 73-84
- Modi V J, Fernando M S U K, Yokomizo T 1990a Drag reduction of bluff bodies through moving surface boundary layer control. *28th Aerospace Sciences Meeting*, Reno, Nevada, Paper No. AIAA-90-0298
- Modi V J, Fernando M S U K, Yokomizo T 1991a Moving surface boundary-layer control: Studies with bluff bodies and application. *AIAA J.* 29: 1400-1406.
- Modi V J, Fernando M S U K, Yokomizo T 1991b Moving surface boundary-layer control as applied to two and three dimensional bodies. *J. Wind Eng. Ind. Aerodyn.* 38: 83-92
- Modi V J, Mokhtarian F 1985 Joukowsky airfoil with circulation control. *AIAA 12th Atmospheric Flight Mechanics Conference*, Snowmass, Colorado, Paper No. 85-1772-CP
- Modi V J, Mokhtarian F, Yokomizo T 1988 Effect of moving surfaces on the airfoil boundary-layer control. *Proc. AIAA Atmospheric Flight Mechanics Conference* (ed.) J E Benek (New York: AIAA) paper. no. AIAA-88-4337
- Modi V J, Mokhtarian F, Yokomizo T, Ohta G, Oinuma T 1987a Bound vortex boundary-layer control with application to v/STOL airplanes. *Proceedings of the IUTAM Symposium on Fundamental Aspects of Vortex Motion* Tokyo, Japan, pp. 196-199
- Modi V J, Mokhtarian F, Yokomizo T, Ohta G, Oinuma T 1987b *Vortex motion* (eds) H Hasimoto, T Kambe (Amsterdam: North-Holland)
- Modi V J, Slater J E 1991 Unsteady aerodynamics and vortex induced aeroelastic instability of a structural angle section. *Proceedings of the 13th Biennial ASME Conference on Mechanical Vibration and Noise* (eds) T C Huang et al (New York: ASME) DE-vol. 37, pp. 177-188
- Modi V J, Swinton P G, McMillan K, Lake P, Mullins D, Akutsu T 1979 On the boundary layer control using the concept of moving surface. *Proceedings of CANCAM 79* (eds) F Ellyin, K W Neale, vol. 2, pp. 621-622

- Modi V J, Swinton P G, McMillan K, Lake P, Mullins D, Akutsu T 1980 Moving surface boundary layer control for aircraft operation at high incidence. *Proceedings of the AIAA 7th Atmospheric Flight Mechanics Conference*, Danvers, MA, Paper No. 80-1621, pp. 515-522
- Modi V J, Swinton P G, McMillan K, Lake P, Mullins D, Akutsu T 1981 *J. Aircraft* 18: 963-968
- Modi V J, Ying B, Yokomizo T 1990b Boundary-layer control of bluff bodies through momentum injection. *SAE International Bus/Truck Meeting and Exposition* Detroit, Paper No. 90-2225
- Modi V J, Ying B, Yokomizo T 1990c *SAE Trans.*, *J. Commercial Vehicles* 99: 778-794
- Modi V J, Ying B, Yokomizo T 1991c An approach to design of the next generation of fuel efficient trucks through aerodynamic drag reduction. *Proceedings of the ASME Winter Annual Meeting* (eds) S A Velinsky, R M Fries, I Haque, D Wang (New York: ASME) DE-vol. 40, pp. 465-482
- Modi V J, Ying B, Yokomizo T 1991d Effect of momentum injection on the aerodynamics of several bluff bodies. *Proceedings of the 8th International Conference on Wind Engineering*, (ed.) A G Davenport (in press)
- Modi V J, Yokomizo T 1992 On the boundary-layer control through momentum injection: Numerical flow visualization and experimental studies. *Proceedings of the 6th International Symposium on Flow Visualization* (eds) Y Tanida, H Miyashiro (Berlin: Springer-Verlag)
- Mokhtarian F 1988 *Fluid dynamics of airfoils with moving surface boundary-layer control*. Ph D thesis, University of British Columbia, Vancouver
- Mokhtarian F, Modi V J 1984 Fluid dynamics of airfoils with circulation control for V/STOL application. *Proc. AIAA 11th Atmospheric Flight Mechanics Conference* (ed.) L B Schiff (New York: AIAA) paper no. AIAA-84-2090
- Mokhtarian F, Modi V J 1986 Fluid dynamics of airfoils with moving surface boundary-layer control. *Proc. AIAA Atmospheric Flight Mechanics Conference* (New York: AIAA) paper no. AIAA-86-2184 CP
- Mokhtarian F, Modi V J 1988 Fluid dynamics of airfoils with moving surface boundary layer control. *J. Aircraft* 25: 163-169
- Mokhtarian F, Modi V J, Yokomizo T 1988 Rotating air scoop as airfoil boundary-layer control. *J. Aircraft* 25: 973-975
- National Research Council 1966 Report of the Research Co-ordination Group on Boundary Layer Control to Suppress Separation, Associate Committee on Aerodynamics
- Rosenhead L 1966 *Laminar boundary layers* (Oxford: University Press)
- Schlichting H 1968 *Boundary-layer theory* (New York: McGraw-Hill)
- Sovaran G, Morel T, Mason T W Jr 1978 Aerodynamic drag mechanisms of bluff bodies and road vehicles. *Proceedings of the Symposium held at the General Motors Research Laboratories* (New York: Plenum)
- Swanson W M 1961 The magnus effect: A summary of investigation to date. *Trans. ASME, Basic Eng.* 83: 461-470
- Thwaites B 1960 *Incompressible aerodynamics* (London: Clarendon)
- Wacker T 1985 *A preliminary study of configuration effects on the drag of a tractor-trailer combination*. MASc thesis, University of British Columbia, Vancouver
- Welt E 1988 *A study of nutation dampers with application to wind-induced oscillations*, Ph D thesis, University of British Columbia, Vancouver
- Ying B 1991 *Boundary-layer control of bluff bodies with application to drag reduction of tractor-trailer truck configurations*. MASc thesis, University of British Columbia, Vancouver
- Zdravkovich M M 1980 Review and assessment of effectiveness of various aero and hydro-mechanic means for suppressing vortex shedding. *Proceedings of the 4th Colloquium on Industrial Aerodynamics* (eds) C Kramer et al (Aachen, Germany: Fachhochschule) part 2, pp. 29-46

## Design and analysis trends of helicopter rotor systems

INDERJIT CHOPRA

Center for Rotorcraft Education and Research, Department of Aerospace Engineering, University of Maryland, College Park, Maryland 20742, USA

**Abstract.** To overcome many of the problems associated with conventional articulated rotor systems, new rotor systems are being contemplated. In this paper, the state-of-art technology of advanced rotor systems is assessed. Advanced rotors include hingeless, bearingless, composite, circulation control, tilt and advanced geometry rotors. The paper reviews mathematical modelling, analysis methods, past and recent developments, potential limitations and future research needs in each system. Also, the potential of incorporation of structural optimization methodology and smart structures technology in rotors to improve the efficiency and capabilities of rotorcraft is discussed.

**Keywords.** Articulated rotor systems; structural optimization; rotorcraft; helicopter rotor systems.

### 1. Introduction

Conventional articulated rotor systems, routinely used in current helicopters, suffer from problems that include high vibratory loads, susceptibility to ground resonance instability, low control power, high operating cost and poor performance in high speed and high load conditions. To improve some of these deficiencies, and also to further expand the flight missions of military and civilian helicopters, many new rotor systems are being contemplated. Examples of new rotor systems are: hingeless, bearingless, composite, circulation control, tilt and advanced geometry rotors. Currently, there are numerous drawbacks of these advanced rotor systems that include inadequate analytical tools, new and sometimes severe dynamic problems, insufficient test and flight experience, and enormous development cost. An assessment of the state-of-the-art of various advanced rotor systems, including past, present and future developmental plans and analysis techniques, are presented. Further, the applications of structural optimization and smart structures technology to the rotor system are discussed in order to reduce vibration, increase aeromechanical stability, minimize blade stresses and improve basic performance.

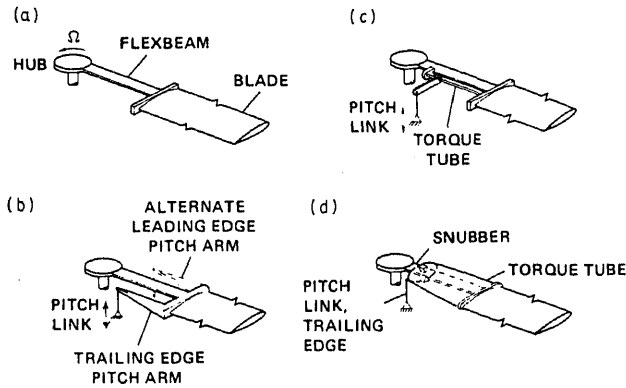
Rotor aeroelastic analyses in forward flight normally consist of calculating vehicle trim, steady response and stability of perturbation motion. Trim involves the calculation of control positions and vehicle orientation for a prescribed flight condition. Typically, there are two types of trim, free-flight and wind-tunnel trim.

The former simulates the free-flight propulsive condition, and the latter simulates the test conditions in the wind tunnel. Steady blade response involves the calculation of blade deflections around the azimuth for one complete revolution. Using trim controls, steady response is calculated from nonlinear blade equations containing periodic terms as a forced response problem, using either the harmonic balance method, a time integration technique, or a finite element in time approach. For stability solutions, the perturbation equations of motion are linearized about the steady response and solved for stability roots. The rotor-body linearized equations are typically transformed into a fixed reference frame (body frame). These contain periodic terms in forward flight and are solved for roots using either Floquet transition matrix theory or constant coefficient approximation. All these phases of analysis are inherently coupled; however to simplify the analyses, these are generally uncoupled and solved individually. Then, coupling between different analyses is achieved through an iterative process. Invariably all analyses assume that all blades are identical and exposed to an identical environment (tracked condition). Chopra (1990) discusses the state of art of different analysis schemes to solve rotor dynamics problems. With new rotor systems, different phases of analyses need to be reformulated or modified and these are discussed in this paper.

## **2. Hingeless and bearingless rotors**

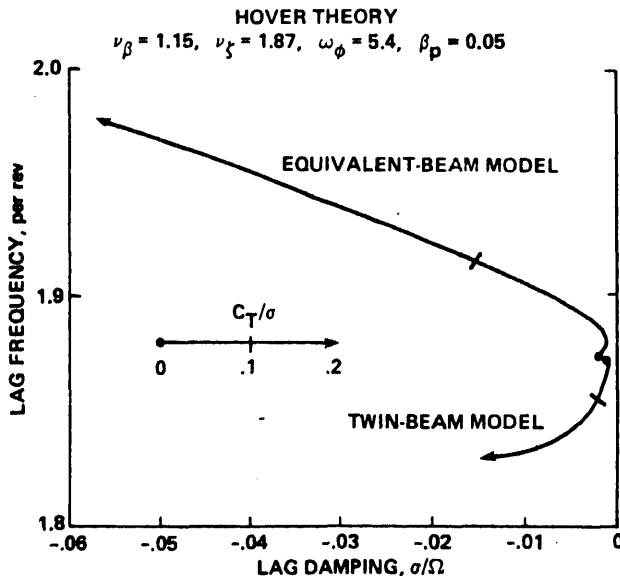
Current rotor designs tend towards hinge- and bearingless rotors because of reduced costs and maintenance (fewer parts), better hub design (simple and clean aerodynamically), and superior handling qualities. In a hingeless rotor, flap and lag hinges are eliminated, and the pitch bearing is eliminated as well in a bearingless rotor. Such rotors are now becoming feasible because of the emergence of composite technology. Because these rotors are stiffer than articulated rotors, they are exposed to higher dynamic stresses. Due to stress and weight considerations, hingeless and bearingless rotors are designed as soft-inplane rotors which make them susceptible to aeromechanical instability. Articulated rotors are protected from ground resonance instability by mechanical lag dampers. However, the effectiveness of mechanical lag dampers is reduced for hingeless and bearingless rotors because of small lag displacements near the root. Also, with bearingless rotors, there is a redundancy of load paths at the root and many of the currently available dynamic codes cannot handle this. The increased forward speed and maneuverability expected out of future helicopters further aggravates this problem. It is therefore a challenging task to achieve aeromechanical stability of hingeless and bearingless rotors.

The distinguishing feature of a bearingless rotor is a torsionally soft flexbeam located between the main blade and hub. Pitch control to the blade is applied through a torsionally stiff torque tube by rotating it with pitch links, which in turn twists the flexbeam (figure 1). Large deflections in the flexbeam, especially in torsion, result in nonlinear bending-torsion couplings. Unlike the analysis of hingeless and articulated rotors where axial deflection can be routinely eliminated, analysis of bearingless rotors becomes complicated because axial deflections are needed to determine leads in different beams. Also, since the blade pitch is determined iteratively by adjusting the position of pitch links, i.e., boundary conditions at the torque tube, the analysis of a bearingless rotor is quite involved, especially in forward flight. Early analyses adopted an "equivalent beam approach" where flexbeams and torque tube together

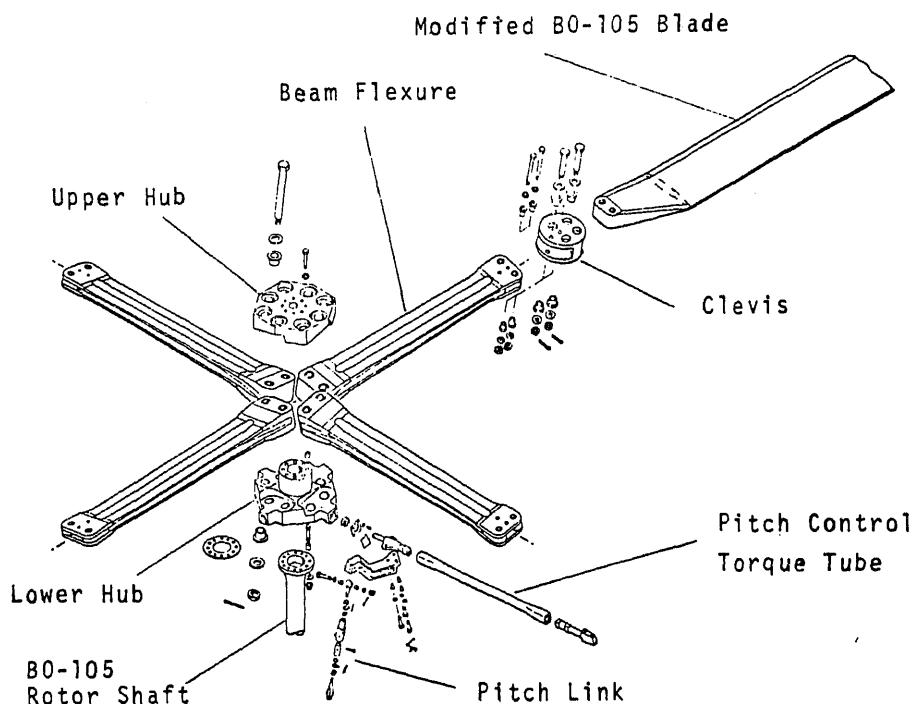


**Figure 1.** Principal configurations for bearingless rotor systems. (a) No pitch-control system; (b) cantilever pitch arm; (c) flexible torque tube; (d) torque tube and snubber.

were approximated as a single beam whose stiffness characteristics had to be calculated for each flight condition and then analysed using a hingeless rotor analysis. It was shown by Sivaneri & Chopra (1984) that equivalent beam modelling cannot accurately simulate nonlinear structural couplings and can lead to erroneous stability results (figure 2). Finite element analysis appears natural to analyse bearingless rotors. It was applied the first time by Sivaneri & Chopra (1984) to calculate the hover flap-lag stability of a simple bearingless configuration, and later on expanded (Dull & Chopra 1988; Jang & Chopra 1988; Jang & Chopra 1989; Wang *et al* 1990), to analyse the aeromechanical stability of several bearingless configurations in hover and forward flight. Each of the flexbeams, torque tube and main blade are discretized into beam elements, and then certain displacement compatibility relations at the clevis (where flexbeams, torque tube and main blade join) are introduced into the assembled matrices. For hovering flight, the steady deflected position of the blade is calculated directly from nonlinear finite element equations using pitch links position as the boundary condition. The desired collective pitch is achieved iteratively by adjusting



**Figure 2.** Bearingless rotor stability results (Shivaneri & Chopra 1984).



**Figure 3.** Boeing BO 105/BMR bearingless rotor system.

the pitch links position. In forward flight, the assembled finite element equations are nonlinear and periodic and also, the pitch link position varies along the azimuth. To reduce the computation time, these are first reduced to the modal space as a few (about six) normal mode equations. Then these are further discretized into nonlinear algebraic equations using finite-elements in time and solved iteratively. Other bearingless rotor analyses are discussed by Chopra (1990).

The Army's Aeroflightdynamics Directorate at Ames performed extensive aero-mechanical stability testing in hover on hingeless and bearingless rotor models. These data form a part of the ITR data set (McNulty & Bousman 1983) and are widely used by researchers to validate their analyses. Boeing built the first full-scale bearingless rotor (BO 105/BMR) and tested it successfully in the NASA Ames 40 × 80 ft wind tunnel. It consisted of twin flexbeams (C-beams) and a single torque tube rod and did not have lag dampers (figure 3). It was found to be marginally stable for many of the flight conditions. Also, it was flight tested successfully. Boeing, under the ITR program, also built a Froude-scaled four-bladed (diameter 6 ft) bearingless rotor (figure 4) which was extensively tested at Maryland's Glenn L Martin wind tunnel (Wang *et al* 1989; Wang & Chopra 1990). It consists of a single flexbeam with a wrap-around torque tube with a vertical offset of the cuff snubber attachment point. The shear pin, in the form of a shaft-mounted spherical pivot, restrains the in-plane and out-of-plane motion of the cuff and introduces a negative pitch-lag coupling. This means that the lagwise shear reaction, coupled with vertical offset of the pivot from the elastic axis, produces feathering motion due to lag motion. Reducing pitchlink stiffness increases the effectiveness of the cuff-restraint and causes a stabilizing influence on air resonance stability (figures 5 & 6). McDonnell Douglas and Bell respectively built HARP (figure 7)

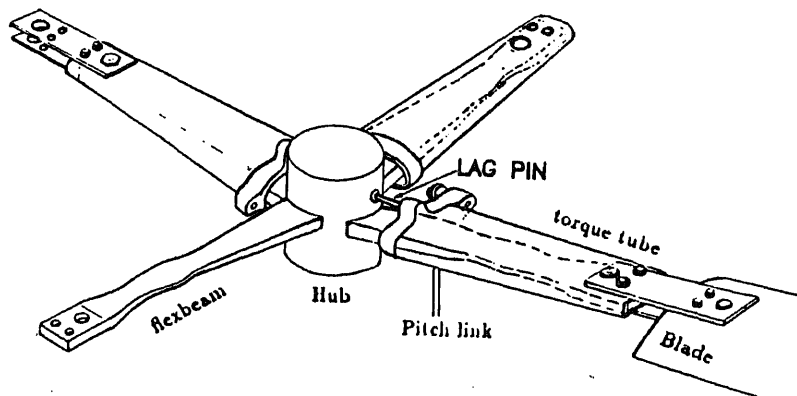


Figure 4. Boeing ITR bearingless rotor model.

and Model 680 (figure 8) full-scale bearingless rotors. In both these rotors, rotary type elastomeric lag dampers are installed between the torque tube and flexbeam. Also in the HARP rotor, the pitch link is slightly inclined to introduce negative pitch-lag coupling. There are several production bearingless tail rotors, such as on the Sikorsky Black Hawk (UH-60A) and S-76, MBB BO 105 and BK 117, and McDonnell Douglas Apache (AH-64A).

A key element in the design of a bearingless rotor is the flexbeam. To reduce hub drag and weight, it needs to be light and compact. The flexbeam undergoes large elastic deflections, say  $\pm 5^\circ$  of flap deflection and  $\pm 15^\circ$  of twist. Therefore, the flexbeam is designed for centrifugal force and large elastic deflections (bending stresses). It is necessary for the blade lag frequency to be as high as possible, but a lag frequency greater than 0.7/rev will lead to high dynamic stresses. The cross-section of the flexbeam is tailored along its length such that the flapping flexure (virtual hinge) falls inboard of lag and torsion flexure. It is now well established that soft-inplane hingeless and bearingless rotors are susceptible to air resonance instability, becoming worse at high thrust levels and Lock numbers (Chen & Chopra 1991, figures 9a and 9b). To stabilize the rotor-body system, it is necessary to include elastomeric lag dampers and/or negative pitch-lag coupling. The characteristics of elastomeric dampers are

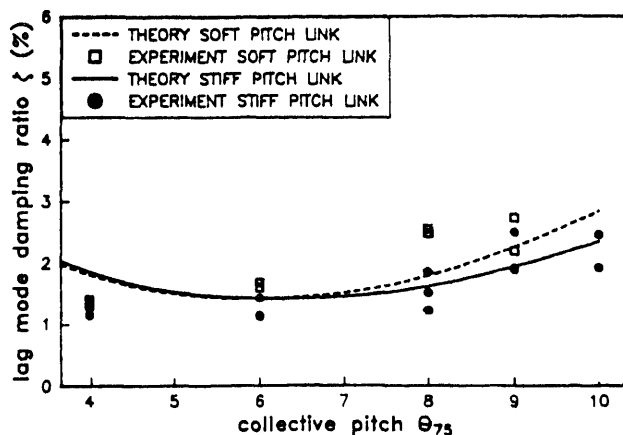


Figure 5. Aeromechanical stability ITR/BMR,  $v_\beta = 1.1/\text{rev}$ ,  $v_\zeta = 0.7/\text{rev}$ . Lag mode stability at  $\mu = 0.35$ , forward shaft tilt =  $4^\circ$  (Wang & Chopra 1990).



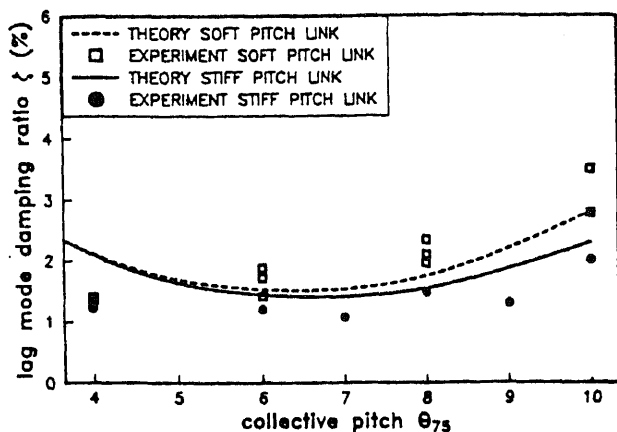


Figure 6. Lag mode stability at  $\mu=0.35$ , forward shaft tilt =  $8^\circ$  (Wang & Chopra 1990).

highly nonlinear functions of lag displacement and velocity and contribute to both stiffness and damping of lag mode. As shown in Chen & Chopra (1991), unless the nonlinear characteristics of elastomeric dampers are modelled properly, it can lead to non-conservative results (figure 10). As far as pitch-lag coupling is concerned, one has to be careful, a negative value stabilizes air resonance stability but destabilizes ground resonance stability. There is no doubt that most of the future rotors will be built as bearingless rotors. As an example, the US Army's RAH-66 Comanche helicopter will incorporate a bearingless rotor. It is expected that more data will be available in future on loads, stability and response of bearingless rotor, which will help in validation of analyses and in our understanding of these systems.

### 3. Composite rotors

Advanced composites are poised for a quantum leap in the rotorcraft industry because of their superior fatigue characteristics as compared to metals, their higher stiffness-

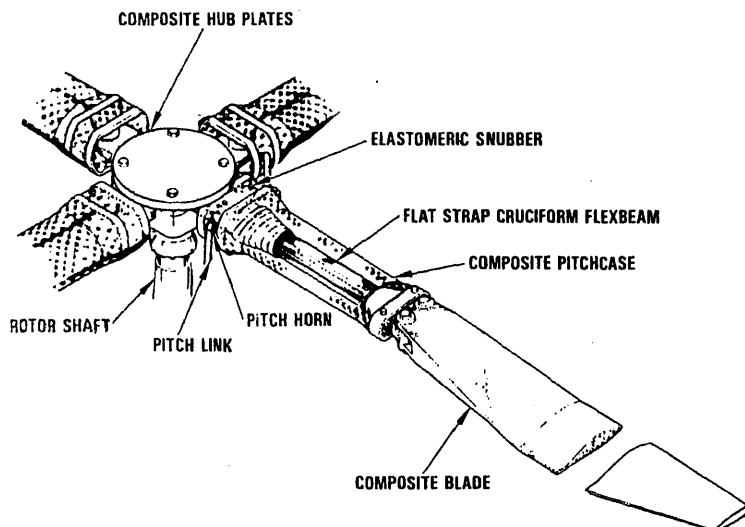
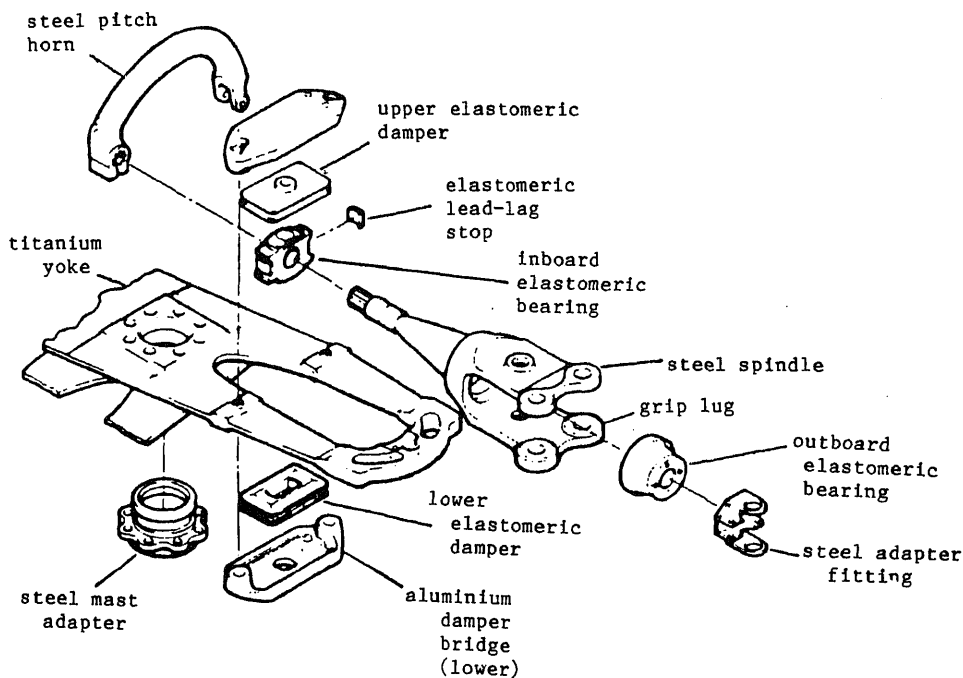


Figure 7. McDonnell Douglas HARP bearingless rotor.



**Figure 8.** Bell Model 680 rotor hub configuration.

to-weight properties, their flexibility in tailoring structural characteristics, and their potential to lower production and operating costs. The rotorcraft industry is moving vigorously to apply composite technology to the construction of rotor systems and airframe structures. For example, the Army's ACAP systems, McDonnell Douglas's HARP bearingless rotor, Bell's Model 680 bearingless rotor, Boeing's Model 500 airframe, and Sikorsky's X-Wing were built extensively out of advanced composites.

There are also many concerns about composites, including lack of understanding of the structural couplings, failure mechanisms, and susceptibility to moisture and lightning strikes. At this time, an extreme level of conservatism is used in rotor designs with composites and thus the potential benefits of their structural couplings are not exploited fully. It is now becoming clear that for modelling of composite blades, nonclassical effects such as section warping, transverse shear and in-plane elasticity become quite important and need to be introduced in the analysis. With the availability of reliable analysis tools for a composite blade, it will be possible to tailor structural properties of the blade to reduce vibration, improve performance and increase aeromechanical stability. The state of art on analysis of a composite blade with a general ply layup is not available at this time.

An important load carrying member of the blade, i.e., spar, is generally a thin-walled closed-section tube (figure 11). Therefore, much can be learned about the behaviour of a composite blade from the analysis of a thin-walled single-cell slender beam. Accordingly, many formulations have been developed recently to model composite beams. These range from simple analytical models (H Minguet & Dugundji 1990; Rehfield *et al* 1990; Smith & Chopra finite elements models (Bauchau & Hong 1988; Stemple & Lee 1989; Friedmann 1989). Also, there have been some selected validation:

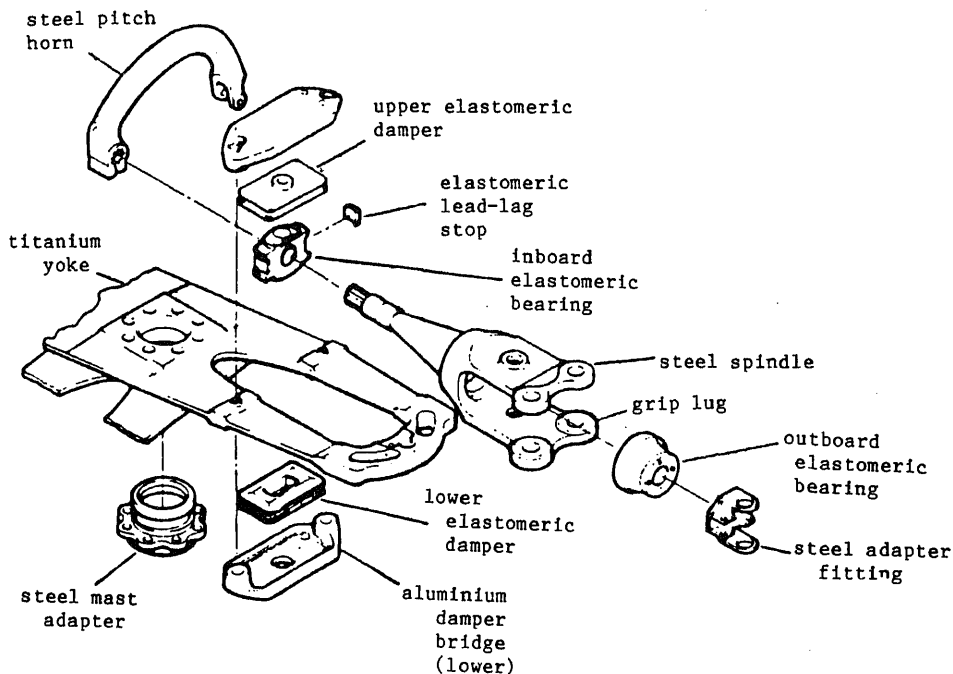
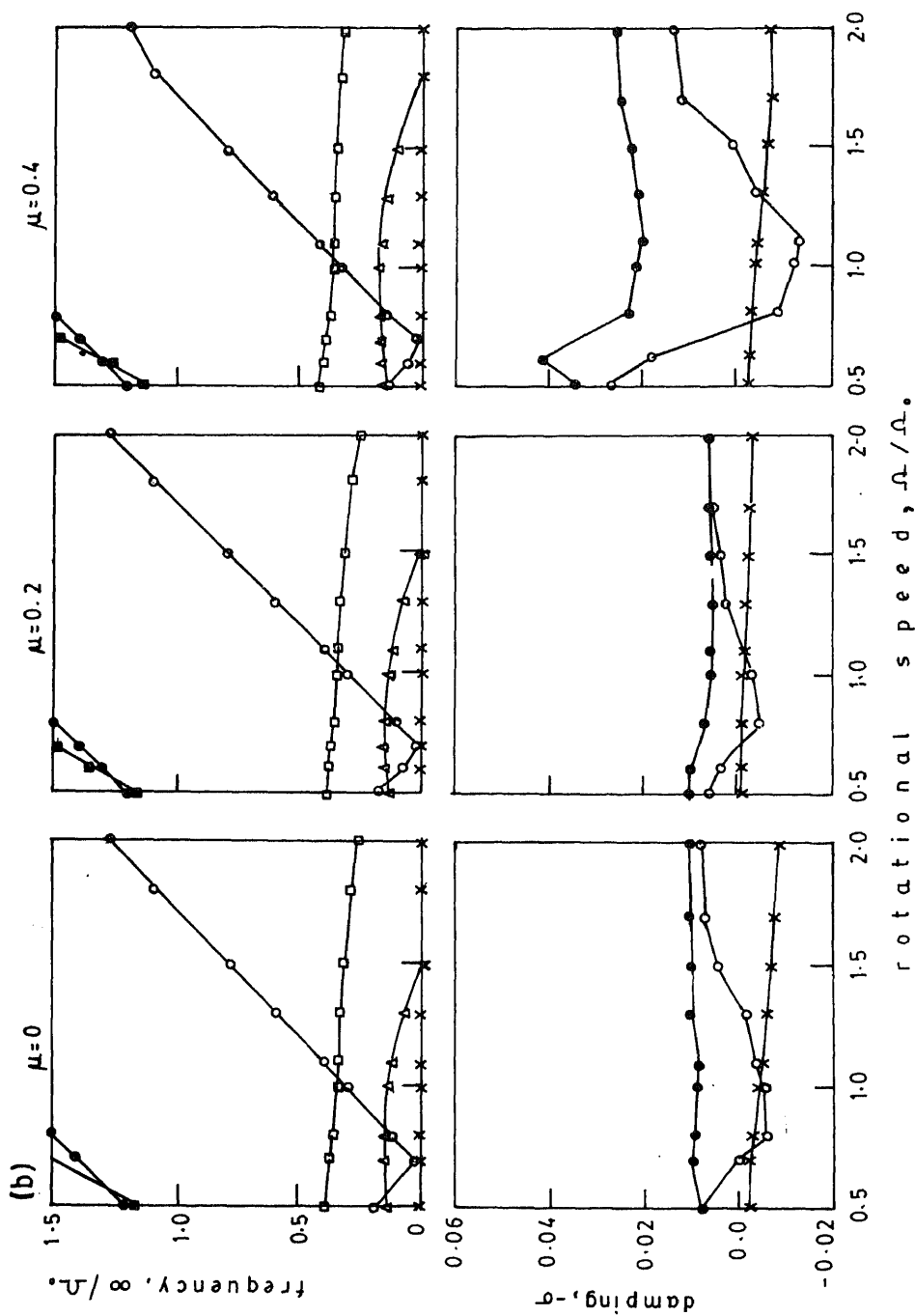


Figure 8. Bell Model 680 rotor hub configuration.

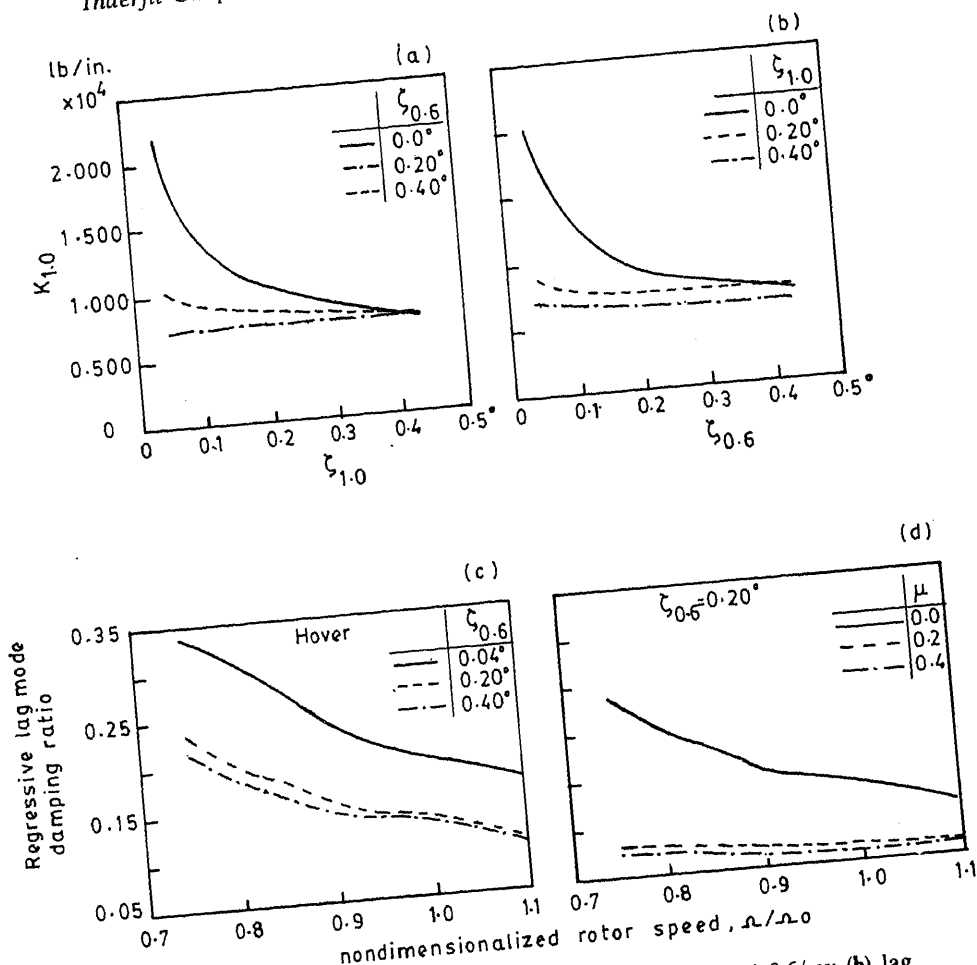
to-weight properties, their flexibility in tailoring structural characteristics, and their potential to lower production and operating costs. The rotorcraft industry is moving vigorously to apply composite technology to the construction of rotor systems and airframe structures. For example, the Army's ACAP systems, McDonnell Douglas's HARP bearingless rotor, Bell's Model 680 bearingless rotor, Boeing's Model 500 airframe, and Sikorsky's X-Wing were built extensively out of advanced composites.

There are also many concerns about composites, including lack of understanding of the structural couplings, failure mechanisms, and susceptibility to moisture and lightning strikes. At this time, an extreme level of conservatism is used in rotor designs with composites and thus the potential benefits of their structural couplings are not exploited fully. It is now becoming clear that for modelling of composite blades, nonclassical effects such as section warping, transverse shear and in-plane elasticity become quite important and need to be introduced in the analysis. With the availability of reliable analysis tools for a composite blade, it will be possible to tailor structural properties of the blade to reduce vibration, improve performance and increase aeromechanical stability. The state of art on analysis of a composite blade with a general ply layout is not available at this time.

An important load carrying member of the blade, i.e., spar, is generally a thin-walled closed-section tube (figure 11). Therefore, much can be learned about the behaviour of a composite blade from the analysis of a thin-walled single-cell slender beam. Accordingly, many formulations have been developed recently to model thin-walled composite beams. These range from simple analytical models (Hodges *et al* 1989; Minguet & Dugundji 1990; Rehfield *et al* 1990; Smith & Chopra 1991) to detailed finite elements models (Bauchau & Hong 1988; Stemple & Lee 1988; Kosmatka & Friedmann 1989). Also, there have been some selected validations for these models



**Figure 9.** (b) Effect of forward speed on frequency and damping versus rotational speed for a hingeless rotor,  $\beta - \zeta - \theta - \phi$  model,  $C_T/\sigma = 0.1$ ,  $\gamma = 1.0$ ,  $R_S = 1.0$  (Chen & Chopra 1991). Notations as in figure 9a.



**Figure 10.** Variation of elastomeric stiffness with 1/rev (a) and 0.6/rev (b) lag amplitude (Bir & Chopra 1981). Damping of the regressive lag mode as a function (c) of 0.6/rev lag motion amplitude ( $\mu = 0$ ,  $C_T = 0.08$ ), and (d) of advance ratio ( $C_T = 0.08$ ,  $\zeta = 0.20^\circ$ ) (Bir & Chopra 1981).

for static and vibration characteristics (Nixon 1989; Chandra *et al* 1990; Chandra & Chopra 1992a).

For convenience, composite beams are categorized into symmetric and antisymmetric configurations based on their ply layouts. In a symmetric configuration, the ply layouts on opposite flanges are identical (mirror image) with respect to the beam axis whereas in an antisymmetric configuration, the ply layouts on opposite flanges are of reversed orientation (figure 12). Following the simple linear analysis of Smith & Chopra (1991) for bending and torsion of thin-walled symmetric composite beams:

$$\begin{Bmatrix} M_f \\ M_c \\ T \end{Bmatrix} = \begin{bmatrix} EI_f & 0 & K_{P_s}^f \\ 0 & EI_c & K_{P_s}^c \\ K_{P_s}^f & K_{P_s}^c & GJ \end{bmatrix} \begin{Bmatrix} w'' - \gamma'_{xy} \\ v'' - \gamma'_{xz} \\ \phi' \end{Bmatrix},$$

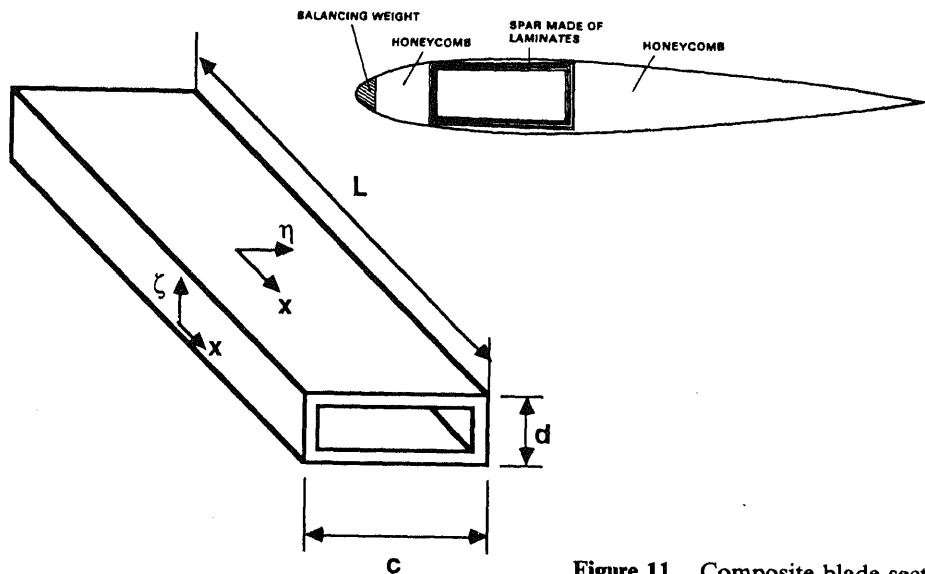
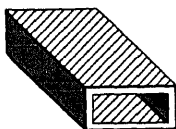


Figure 11. Composite blade section.

where  $M_f$ ,  $M_c$  and  $T$  are respectively flap bending, chordwise bending and torque at a given station, and  $EI_f$ ,  $EI_c$ , and  $GJ$  are respectively effective flap stiffness, chordwise stiffness, and torsional stiffness, and  $w''$  is the flap bending curvature,  $v''$  is the chordwise bending curvature,  $\phi'$  is the twist derivative, and  $\gamma'_{xy}$  and  $\gamma'_{yx}$  are derivatives of cross-section transverse shear strains. This symmetric layup configuration displays bending-torsion coupling and extension-shear coupling. The quantity  $K'_{ps}$  is the flap bending-twisting structural coupling and is quite similar to the classical pitch-flap coupling. This coupling can be achieved through nonzero ply angles on top and bottom laminates. This coupling was seen to have a considerable effect on flap

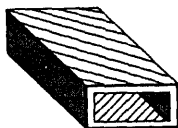
#### Cross-Ply Layup Beam

- Alternating Piles at 0 and 90 deg
- No Elastic Couplings



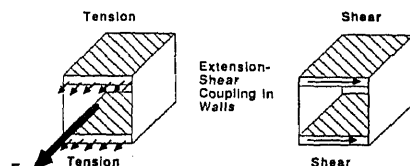
#### Symmetric Layup Beam

- Bending-Torsion Coupling
- Extension-Shear Coupling

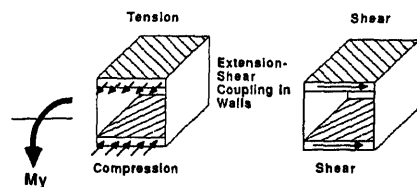


#### Anti-Symmetric Layup Beam

- Extension-Torsion Coupling
- Bending-Shear Coupling

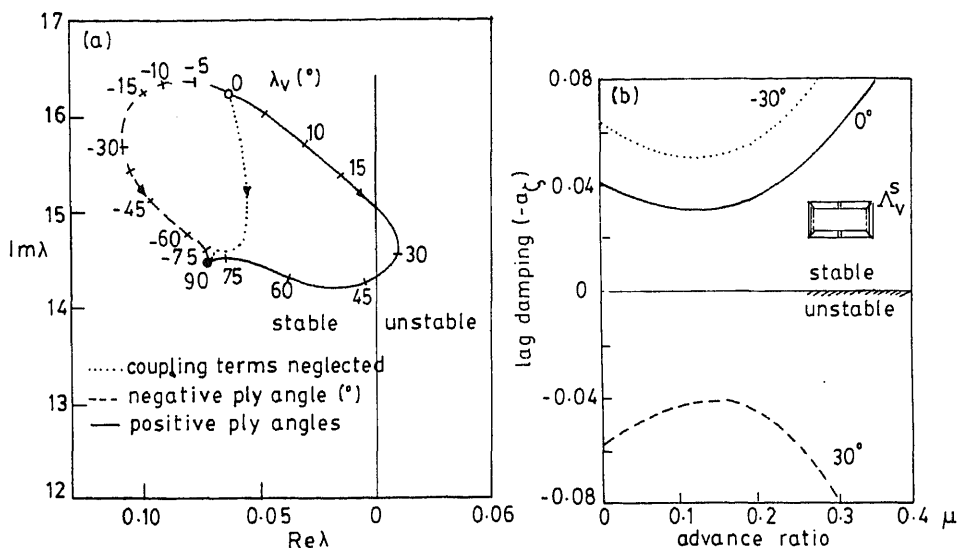


Extension-shear coupling in symmetric layup beams



Bending-shear coupling in anti-symmetric layup beams

Figure 12. Composite layup designation.



**Figure 13.** (a) Root locus for lag mode of a composite rotor (symmetric layup) in hover,  $C_T/\sigma = 0.1$  (Hong & Chopra 1985). (b) Damping of low frequency lag mode of a composite rotor (symmetric layup),  $C_T/\sigma = 0.07$  (Panda & Chopra 1987).

dynamics, and a greater effect on vibration and bending stresses than aeroelastic stability (Hong & Chopra 1985; Nixon 1989; Minguet & Dugundji 1990). The quantity  $K_{p_z}^c$  is the chordwise bending-twisting coupling and is quite similar to classical pitch-lag coupling. This coupling can be achieved through nonzero ply angles on the side laminates or on the vertical web (I-beam). This coupling was seen to have a large effect on aeroelastic stability (Hong & Chopra 1985, 1986; Panda & Chopra 1987) (figure 13).

In a similar way, following the analysis of Smith & Chopra (1991) a simple linear analysis for extension and torsion of thin-walled antisymmetric composite beams is derived as

$$\begin{Bmatrix} T \\ F \end{Bmatrix} = \begin{bmatrix} GJ & K_{p_a} \\ K_{p_a} & EA \end{bmatrix} \begin{Bmatrix} \phi' \\ u' \end{Bmatrix},$$

where  $F$  is the axial force,  $EA$  is the effective extensional stiffness, and  $u'$  is the axial deflection derivative.  $K_{p_a}$  is the extension-twisting structural coupling and is caused by the nonzero ply angles on top and bottom laminates or side laminates. This coupling was seen to have a considerable influence on blade dynamics, including lag mode stability (Hong & Chopra 1985, 1986; Panda & Chopra 1987) (figure 14).

Including effects of transverse shear, causes bending-shear coupling for this configuration (Smith & Chopra 1991).

$$\begin{Bmatrix} M_f \\ Q_y \end{Bmatrix} = \begin{bmatrix} EI_f & K_{p_y}^s \\ K_{p_y}^s & GA_y \end{bmatrix} \begin{Bmatrix} w'' - \gamma'_{xz} \\ \gamma'_{xy} \end{Bmatrix}$$

$$\begin{Bmatrix} M_c \\ Q_z \end{Bmatrix} = \begin{bmatrix} EI_c & K_{p_z}^s \\ K_{p_z}^s & GA_z \end{bmatrix} \begin{Bmatrix} v'' - \gamma'_{xy} \\ \gamma'_{xz} \end{Bmatrix},$$

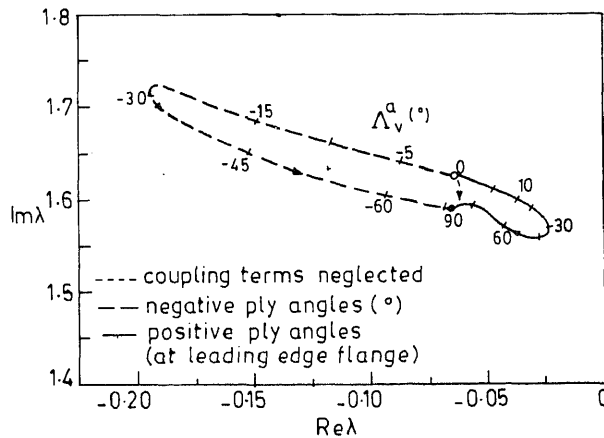


Figure 14. Roots locus for lag mode of a composite rotor (anti-symmetric layup) in hover,  $C_T/\sigma = 0.1$  (Hong & Chopra 1985).

where  $Q_y$  and  $Q_z$  are net shear forces on the cross section,  $GA_y$  and  $GA_z$  are transverse shear stiffnesses and  $K_{P_y}^S$  and  $K_{P_z}^S$  are bending-shear couplings. It is interesting to note that for antisymmetric layup configurations, the extension and torsion of the beam are not elastically coupled to the bending and shearing of the beam.

The analysis of thin-walled closed-section slender composite beams showed that nonclassical effects, such as cross-sectional warping and transverse shear deflections, become important. An improved model for cross-sectional warping was developed (Smith & Chopra 1991) which included the effect of variations in stiffness around the cross-section. As shown in figure 15, the effects of warping can become quite severe even for thin-walled composite beams. With the existence of elastic couplings, the effects of transverse shear can become quite significant even for slender beams. For symmetric layup configurations, extension and transverse shear are coupled, and for antisymmetric layup configurations, bending and transverse shear are coupled. Figure 16 shows considerable reduction of effective bending stiffness due to bending-shear coupling in an antisymmetric layup beam. Another important element of composite beam analysis is the adequate representation of two-dimensional in-plane

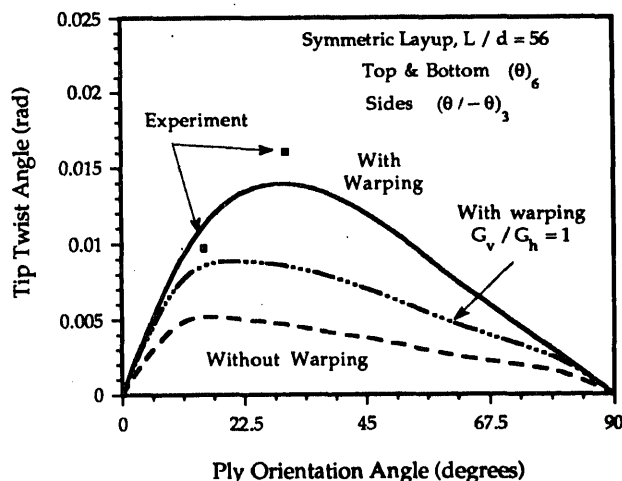
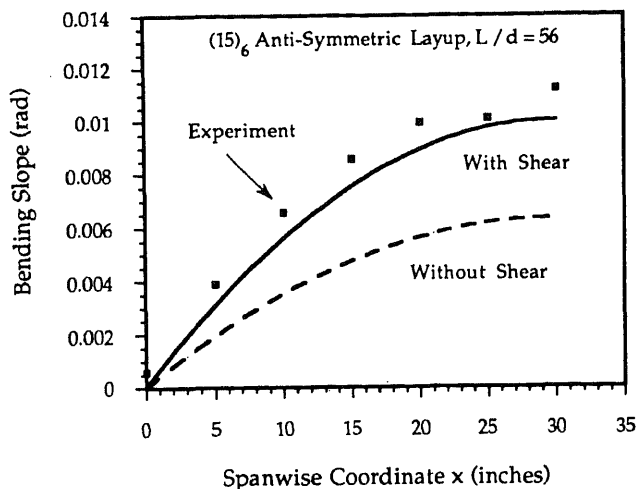


Figure 15. Tip twist under unit tip bending load for a symmetric layup thin-walled composite box beam,  $L = 30$  in.,  $d = 0.537$  in.,  $C = 0.953$  in. (Smith & Chopra 1991).





**Figure 16.** Bending slope under unit tip bending load for an antisymmetric layup thin-walled composite box beam,  $L = 30$  in.,  $d = 0.537$  in.,  $C = 0.953$  in. (Smith & Chopra 1991).

elastic behaviour of plies. Smith & Chopra (1991) consider three different formulations: classical beam theory (plane strain), plane stress condition and no in-plane forces and moments. The third approach shows better correlation with experimental data.

There have been some attempts to verify these structural couplings experimentally. In Chandra *et al* (1990), thin-walled rectangular section beams were built out of graphite/epoxy pre-impregnated tape using an autoclave molding technique and under bending, torsion and extensional loads. Uniform symmetric and antisymmetric layup beams were built for several cases. Measured bending slope and twist distributions were correlated satisfactorily (Chandra *et al* 1990; Smith & Chopra 1991). These beams were also tested for their rotating vibration characteristics in the 10-foot diameter vacuum chamber (Chandra & Chopra 1992a). Measured frequencies and mode shapes correlated satisfactorily with analytical predictions. At this time, there is better understanding of thin-walled closed section slender composite beams under static and dynamic loads than of open-section composite beams.

Open-section composite beams, such as I-beams, are used in the fabrication of flexbeams. Hong & Chopra (1986) showed that bending-twist and extension-twist couplings introduced through ply orientations in I-beams influenced the dynamics of bearingless rotors substantially. The structural model used in that study was based on the solid section approach; constrained warping and transverse shear effects were not considered. Chandra & Chopra (1991) developed a new formulation based on the Vlasov theory to analyse open-section composite beams. Each laminate (web or flange) is separately treated as a two-dimensional plate, and then using geometric considerations, two dimensional stress and displacement fields are related to one dimensional beam forces and moments. Effects of transverse shear were included. In order to validate this analysis, graphite/epoxy and kevlar/epoxy I-beams were fabricated and tested under bending and torsion loads. It was shown that the torsional stiffness of I-beams is significantly influenced by restraining the warping deformations (figure 17). About a 600% increase in torsional stiffness due to constrained warping is noticed for graphite/epoxy beams with a slenderness ratio of 30. Using this new modelling, the calculated response of composite I-beams showed excellent correlation with measured data (figure 18). Rotating vibration characteristics of composite I-beams were determined experimentally in the vacuum chamber and correlated

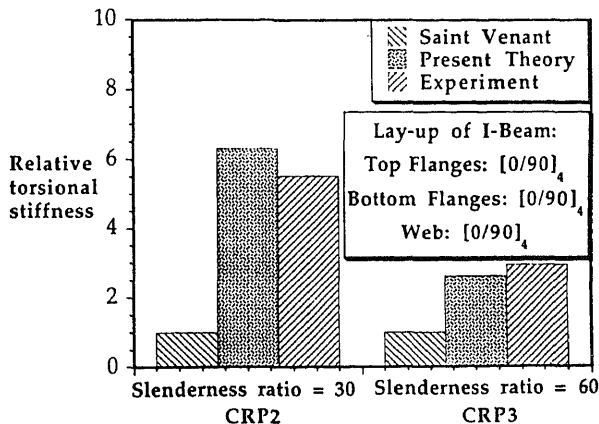


Figure 17. Torsional stiffness of a symmetric graphite-epoxy I-beam under unit tip torsional load, length = 36 in., width = 1 in. (Chandra & Chopra 1991).

successfully with analytical predictions (Chandra & Chopra 1991b). Static and vibration characteristics of multi-cell elastically coupled composite blades were calculated using modified Vlasov theory and compared satisfactorily with measured data (Chandra & Chopra 1992).

Recently, using improved composite beam modelling, comprehensive rotor aeroelastic analysis was reformulated and results were calculated for hub loads, blade response and aeromechanical stability for realistic rotor configurations (Smith & Chopra 1992). A new nineteen degrees of freedom shear flexible element was developed for blade discretization. Again, this study showed that elastic couplings introduced through the composite blade spar have a powerful influence on both shaft-fixed blade stability and rotor-body stability.

By placing plies at specific layup and orientation, the desired structural couplings can be tailored. Therefore, there is great potential for composite tailoring of rotor blades to minimize blade stresses and vibration and to increase aeroelastic stability.

#### 4. Circulation control rotors

A circulation control rotor utilizes circulation control (CC) airfoils. A CC airfoil typically consists of a quasi-elliptical profile with a thin jet of air blown from a spanwise slot near the trailing edge (figure 19). The jet remains attached at the rounded

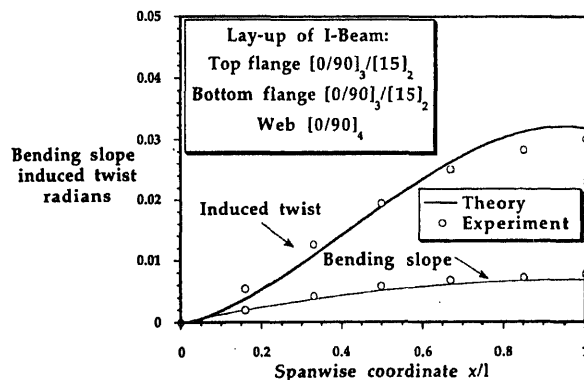


Figure 18. Variation of bending slope and induced twist in a symmetric graphite-epoxy I-beam under unit tip bending load, length = 36 in., width = 1 in., height = 1/2 in., (Chandra & Chopra 1991).

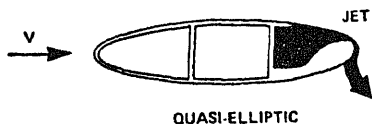


Figure 19. Typical circulation control airfoil.

trailing edge because of the balance of centrifugal force and the suction pressure – called the Coanda effect. At higher blowing, the stagnation point shifts towards the lower surface. With CC airfoils, lift can be controlled by jet momentum as well as by angle of incidence, and high lift coefficients of the order of 3 to 4 can be achieved. In a CC rotor, geometric pitch is held fixed and the cyclic control of lift is achieved through cyclic control of blowing (figure 20). A CC rotor allows a simplified hub design, high thrust at reduced tip speed, and easy implementation of higher harmonic control system. However, the Coanda principle works efficiently only for a limited range of blowing momentum and angle of attack. Also, there are serious concerns of CC technology on rotor dynamics.

The lift on a CC airfoil can be hypothesized into two components: lift due to blowing and lift due to angle of attack. For the blowing-lift, the aerodynamic centre is located near half-chord, whereas for angle-of-attack lift, the aerodynamic centre is located near quarter-chord. Figures 21a and b present lift and moment coefficients for a typical CC airfoil.

The airfoil section lift per unit span is given as

$$L = \frac{1}{2} \rho V^2 c C_l(\alpha, C_\mu),$$

where  $C_\mu$  is the blowing coefficient and is defined as

$$C_\mu = \dot{m} V_j / \frac{1}{2} \rho V^2 c.$$

The quantity  $\dot{m} V_j$  is the jet momentum,  $\frac{1}{2} \rho V^2$  is the dynamic pressure and  $c$  is the chord. With a conventional airfoil, a perturbation of the blade in-plane velocity influences the lift through a change in dynamic pressure and angle-of-attack. With the circulation control airfoil, a perturbation of in-plane velocity has additional effects through the change in momentum coefficient.

$$\begin{aligned} \delta L &= \frac{1}{2} \rho V^2 c \left( \frac{\partial C_l}{\partial \alpha} \delta \alpha + \frac{\partial C_l}{\partial C_\mu} \delta C_\mu + 2 \frac{\partial C_l}{\partial V} \delta V \right) \\ &= \frac{1}{2} \rho V^2 c \left[ \frac{\partial C_l}{\partial \alpha} \delta \alpha + 2 \left( C_l - C_\mu \frac{\partial C_l}{\partial C_\mu} \right) \frac{\delta V}{V} \right]. \end{aligned}$$

#### CIRCULATION CONTROL CONCEPT

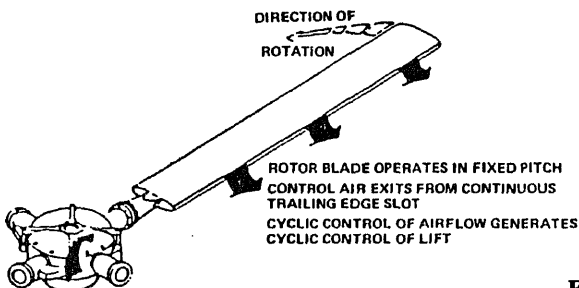


Figure 20. Circulation control rotor.

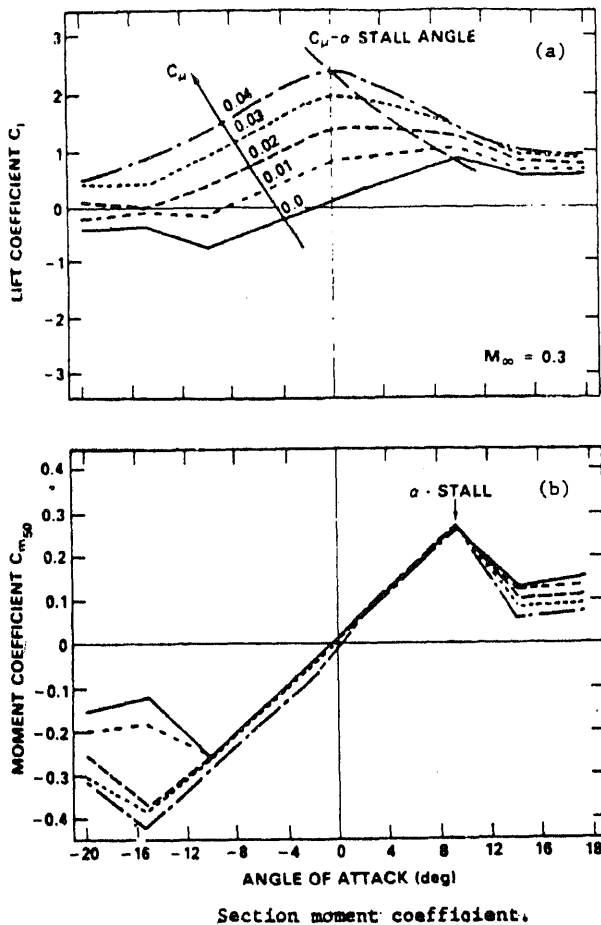


Figure 21. Lift and moment data for a typical CC airfoil: (a) Section lift coefficient and (b) section moment coefficient (Haas & Chopra 1988).

It is assumed that the jet momentum is fixed.

$$\delta C_\mu = \frac{\dot{m} V_j}{\frac{1}{2} \rho c} \left( -\frac{2 \delta V}{V^3} \right) = -2 C_\mu \frac{\delta V}{V}.$$

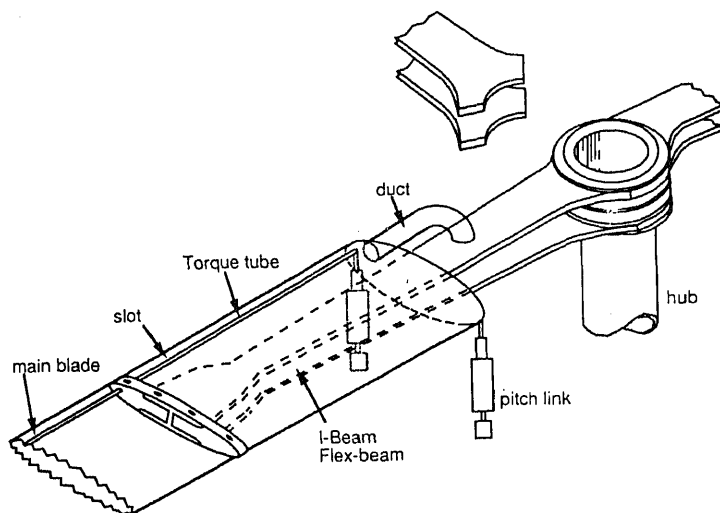
Examining the perturbation of lift, a moderate amount of blowing will reduce the lift perturbation due to a perturbation in in-plane velocity, and a large amount of blowing will alter the sign of lift perturbation. This has considerable influence on blade stability.

Research on CC airfoils including their application to rotors was initiated by Cheeseman (1967) and later on vigorously carried on by David W. Taylor Naval R & D Center over several years (Englar & Applegate 1984). In addition to numerous scaled-model tests in wind tunnels, full-scale rotary-wing and fixed-wing aircraft were also built using CC technology. Kaman built the first full-scale CC rotor, called XH2/CCR (Barnes *et al* 1979). It consisted of four blades with a diameter of 44 ft (13.4 m). It was tested successfully for many flight conditions in the Ames 40 × 80 foot wind tunnel, but the flight program could not proceed because of inadequate cyclic control during transition flight. Lockheed built the first fullscale stoppable rotor, called X-wing

(Reader *et al* 1976). It consisted of four blades with a rotor diameter of 25 ft (7.6 m). An X-Wing aircraft takes off like a helicopter, picks up sufficient forward speed and then the rotor is stopped and it continues flying as a fixed-wing airplane. During the landing flight sequence, the wings are set into rotation and the vehicle transforms into a helicopter. An X-Wing therefore provides the hover and low speed efficiency of a helicopter and the high speed cruise efficiency of a jet airplane. These capabilities are accomplished using dual slot CC airfoils (slots at both leading and trailing edge). Lockheed tested their rotor in the NASA Ames  $40 \times 80$  ft wind tunnel and on the hover stand successfully, but did not attempt flight testing. Later on, Sikorsky built a larger version of the X-Wing for testing on the RSRA aircraft (flying test bed) (Linden & Biggers 1985). It consisted of four blades with a rotor diameter of 56 ft (17.1 m). Because of safety and many other concerns, it was not flight tested. There is no doubt that an X-Wing aircraft is a technology challenge at this time and requires numerous advances and innovations related to CC phenomenon. Blades of the X-Wing are an order of magnitude stiffer than conventional helicopter blades (flap frequency 2.1/rev for X-Wing versus 1.1/rev for hingeless BO 105 rotor) and this aggravates many dynamics problems. To keep dynamic stresses within acceptable limits, the starting and stopping operations of the rotor must be accomplished quickly (say within 10 s). This necessitates an active control of blowing momentum during the change-over phase from rotating blades to fixed-wings and vice versa. Since two of the wings are swept forward ( $45^\circ$ ), aeroelastic divergence can become an issue.

The high lift capability of CC airfoils was investigated to reduce take-off and landing distances of fixed-wing airplanes (STOL application) (Englar *et al* 1978). The Navy is developing a two-bladed stoppable rotor using CC technology for an unmanned aerial vehicle (UAV). Sikorsky examined the potential of CC technology to the application of higher harmonic control of helicopter vibration (Lorber & Carson 1989). Since many harmonics can be incorporated (say up to the first 8), this can be used to actively control blade stresses and augment rotor stability. Circulation control technology has been used successfully on a helicopter tail boom to replace a conventional tail rotor as a means of reacting main rotor torque and providing directional control (NOTAR concept) (Logan 1982).

There are a few selected investigations on the aeroelastic stability of a CC rotor (for a recent review see Chopra 1990). A basic formulation was developed to examine the stability of a CC rotor in hover (Chopra & Johnson 1979). A simple blade model consisting of three degree-of-freedom, flap, lag and feather rotations about hinges, was used, as were airfoil characteristics in the form of analytical expressions. It was shown that trailing edge blowing has substantial influence on blade stability. Contrary to the behaviour of hingeless rotors with conventional airfoils, strong flap-lag instability can occur in a CC rotor at low thrust levels provided blowing momentum is high. Unless some form of kinematic coupling (such as pitch-lag and pitch-flap coupling) is incorporated, structural damping alone may not suffice to correct this instability. Furthermore, because of the location of the elastic axis (spar position) at half-chord, there is a possibility of single-degree torsional instability, showing the need to augment blade torsional damping. From this analysis, it became certain that the CC rotor cannot be designed as soft-in-plane hingeless rotors because of strong aeroelastic instability. Johnson & Chopra (1979) investigated the flight stability of CC rotors in hover using this simple three-degree blade model. Blowing caused a large reduction in the rotor-speed stability derivative. Above a certain blowing level, which depends on the flap frequency and rotor lift, negative speed

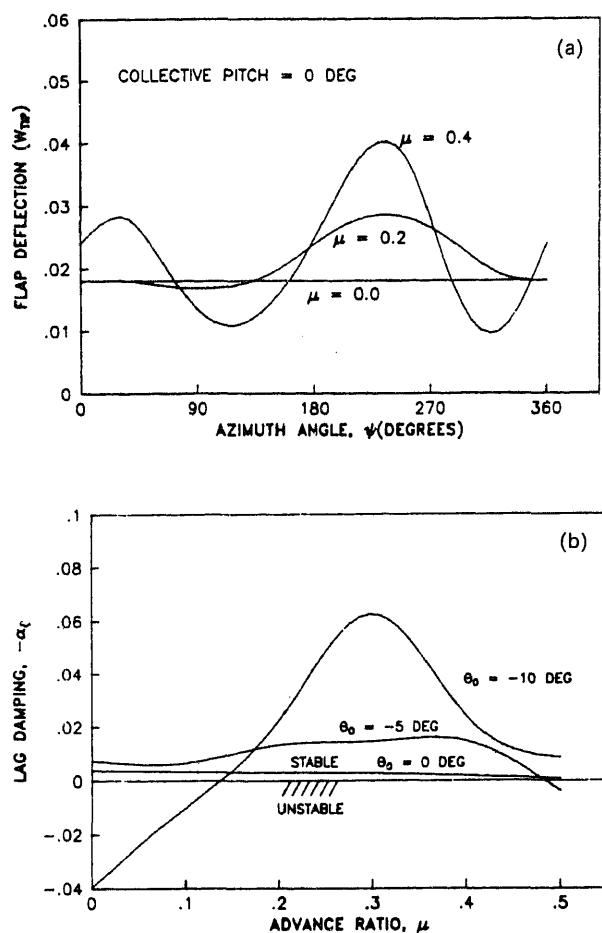


**Figure 22.** Bearingless circulation control rotor.

stability is caused and the dynamic characteristics of the rotor are totally changed. Again, it was shown that the blowing had a substantial influence on the flight dynamics of a helicopter.

The aeroelastic stability of an elastic rotor in hover was investigated (Chopra 1984, 1985) using the finite element formulation for hingeless and bearingless configurations. The airfoil characteristics were taken in the form of data tables. Again, it was shown that trailing edge blowing had an important influence on the aeroelastic stability of the rotor. Negative collective pitch (typically used) caused unstable effects on lag mode stability whereas positive pitch was stabilizing. The dynamics of a bearingless rotor (figure 22) in forward flight was examined (Chopra 1988). This configuration represented the characteristics of Sikorsky's X-Wing rotor. The effects of pneumodynamics and centrifugal pumping (Watkins *et al* 1985) in the pressure duct were included to determine the jet momentum at different stations. In forward flight, the flap response consisted primarily of 2/rev and its amplitude increased with higher speed, showing large bending stresses (figure 23). It is quite clear that some form of active feedback system is required to control blade bending stresses as well as helicopter vibration. Lag mode damping becomes more unstable at low forward speed as collective pitch is decreased (figure 23). Again, this points to the need to augment damping of lag mode either using elastomeric dampers or an active feedback system.

In the above mentioned aeroelastic analyses, quasisteady aerodynamics is used. Raghavan *et al* (1988) formulated a simple model for the unsteady boundary layer for a two-dimensional CC wing. Both unsteady freestream and unsteady blowing were considered. For an elliptic airfoil, an unsteady potential flow solution was calculated using conformal mapping, and the wake was modelled using distributed source panels in the separated region. Substantial unsteady effects were observed for thick airfoils at large blowing rates and large reduced frequencies. Recently, Ghee & Leishman (1990) carried out a systematic experiment to determine the effects of unsteady blowing on the lift development of a two-dimensional wing. The results showed that there was significant increase of lift augmentation over and above the



**Figure 23.** (a) Blade flap bending deflection at tip for a bearingless CC rotor ( $C_T/\sigma = 0.1$ ,  $v_\beta = 2.5/\text{rev}$ ,  $v_c = 2.3/\text{rev}$ ,  $v_\theta = 21/\text{rev}$ ) (Chopra 1988). (b) Effect of collective pitch on low frequency cyclic lag mode (Chopra 1988).

static values due to unsteady blowing. Again, unsteady effects appear important and must be considered for CC rotor analysis.

Haas & Chopra (1988, 1989) showed that blowing has considerable influence on the static divergence and flutter of a CC fixed-wing airplane. It was shown that a loss of blowing control effectiveness occurs on an aft-swept CC wing and may lead to control reversal at high blowing levels. A flutter phenomenon unique to CC wing was discovered, and is referred as "CC flutter". It occurs at low angle of attack with high blowing. It is caused by blowing stall and results in an instability of the low frequency bending mode. This analysis helped to explain the limit-cycle instability that was observed in wind tunnel tests at the United Technology Research Center on the X-wing dynamic model at certain flight conditions. Further, it was shown that with high blowing, the aircraft short period mode also becomes unstable.

There is no doubt that trailing edge blowing has a powerful influence on the aeroelastic response and stability, flight mechanics and performance of fixed-wing as well as rotary-wing aircraft, and must be considered for design development of the system.

## 5. Tilt-rotors

A tilt-rotor aircraft combines the vertical take-off and landing capability of a helicopter and the high-speed cruise capability of a turboprop aircraft (figure 24). The dynamics of a tilt-rotor is more involved than a conventional helicopter rotor because of coupling resulting from wing motion, tilting of the main rotors, high inflow ratio in axial flight, variable rotational speed, and highly twisted blades. In addition to the aeromechanical instabilities of conventional rotors, there are some unique instabilities such as the whirl flutter encountered in a tilt-rotor system. In spite of continuing full-scale development of tilt-rotor aircraft, there are still many concerns about its dynamics.

Over several years, many analytical and experimental studies have been conducted to provide insight into several aeroelastic and dynamic problems associated with a tilt-rotor aircraft. Kvaternik (1973) carried out a systematic proprotor/pylon whirl flutter instability in axial flight. Whirl flutter is caused by coupling of proprotor motion and support motion and occurs at a high forward speed. The analysis was developed for the system undergoing rigid flap motion for blades, three translational and three rotational degrees of motion for the pylon, and vertical bending motion for the wing. Systematic studies were carried out to examine the influence of several design parameters on whirl flutter instability. Calculated results were also correlated successfully with measured stability data obtained from a dynamically scaled semi-span wing/tilt-rotor model in the wind tunnel. The outgrowth of this study was the widely followed code, called PASTA. Johnson (1974) developed a theoretical model for a tilt-rotor on a cantilever wing in high inflow axial flight. The analysis involved first flap and lag bending modes for blades (six degrees for three-bladed rotor), and lowest frequency wing elastic bending (vertical and chordwise) and torsion modes. The pylon with a large mass and moment of inertia is rigidly attached at the tip of the wing. Several parametric studies were conducted to examine the sensitivity of the stability results to various elements in the theoretical model and different proprotor

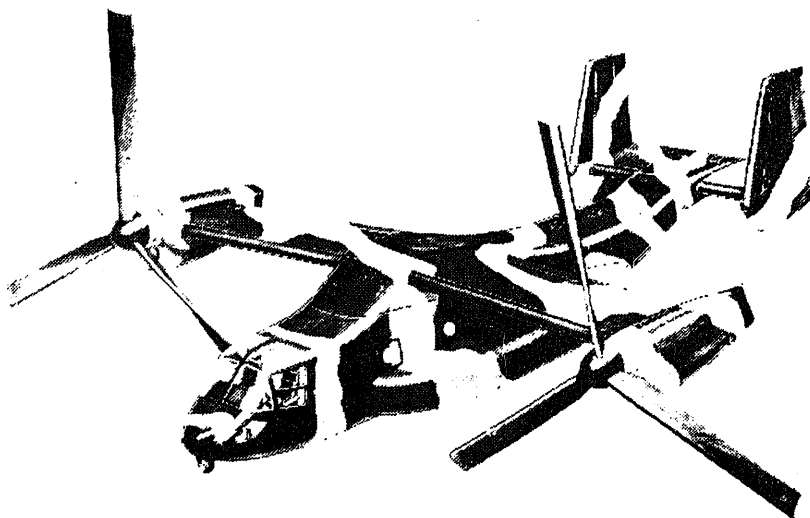


Figure 24. Tilt-rotor aircraft.



configurations. Calculated stability results were correlated successfully with the test results of two full-scale proprotors in the NASA Ames 40- by 80-foot wind tunnel. This study showed the need to include the lag degree of freedom for hingeless proprotor designs. Later on, Johnson (1974b) extended the proprotor dynamic analysis to include elastic coupled rotor modes (flap, lag and torsion) and nonaxial flight. It covered a helicopter in forward flight and a conversion mode flight condition. From the application of the analysis to a gimballed rotor and a hingeless rotor (Johnson 1975), it was shown that for an accurate prediction of dynamic stability, it is important to include blade bending flexibility, a proper representation of root and hub configuration, precise modelling of the rotor support system (i.e. wing), and gimbal and rotor-speed degrees of freedom. These analyses were included in the widely-used code, CAMRAD (Comprehensive Analytical Model of Rotorcraft Aerodynamics and Dynamics) (Johnson 1980, 1981), where in addition to dynamic stability, performance and loads of tilting proprotor aircraft can be also calculated. Bell Helicopter (Popelka *et al* 1985) developed inhouse their own code, called DYN4, to predict the dynamics of a tilt-rotor aircraft. Johnson (1984, 1985) made an assessment of analytical capability for tilt-rotor aircraft in relation to the XV-15 tilt rotor research aircraft. For future methodology development, he recommended the improvement of aerodynamic modelling to include nonuniform inflow and dynamic stall, and a capability to handle new rotor configurations, such as bearingless rotors. He also recommended the acquisition of additional reliable and detailed experimental data to support the development of improved analytical models.

The most critical instability in a tilt-rotor aircraft is whirl flutter. In addition to rotor and wing dynamics, the engine and transmission dynamics have a major influence on this stability. The interconnecting shaft introduces a differential speed mode with a natural frequency of the same order as the wing mode. Tilt-rotor designs (Bell-Boeing JVX and V22) have two heavy engines, one at each wing tip, and it results in several low frequency wing modes which couple with the proprotor modes. In fact, pylon whirl flutter limits the maximum airspeed of tilt-rotor aircraft. One approach to increasing the flutter speed is to design the wing to prevent the vibratory motion from coupling with the proprotor forces. Alternatively, careful design of proprotor aerodynamics can also increase the stability speed. These two approaches, although appearing simple, are difficult to implement because an accurate knowledge of the vibration characteristics of the vehicle is required and these approaches often lead to expensive solutions. A promising approach for increasing the speed of tilt-rotor aircraft is to use an active control system to suppress the coupling between the wing and proprotor. This can be implemented by adjusting the pitch controls of the proprotor based on feedback of wing motion (Vorwald & Chopra 1991). To reduce the influence of external disturbances and measurement noise, the model variables are estimated with a Kalman-Bucy filter. An example showed that using the feedback of vertical wing motion, the stability speed of the Bell XV-15 model could be raised by 20% (figure 25).

In summary, available analyses: lack accurate modelling of complex hubs, including bearingless and elastomeric hubs; cannot model adequately composite couplings; use simple modelling for advanced-tip rotors; use very restrictive aerodynamic representation; and incorporate limited degrees of motion for wings and blades. However, because of the growing interest by NASA in a high speed tilt-rotor aircraft, many new research initiatives are taking place which will help to solve these problems.

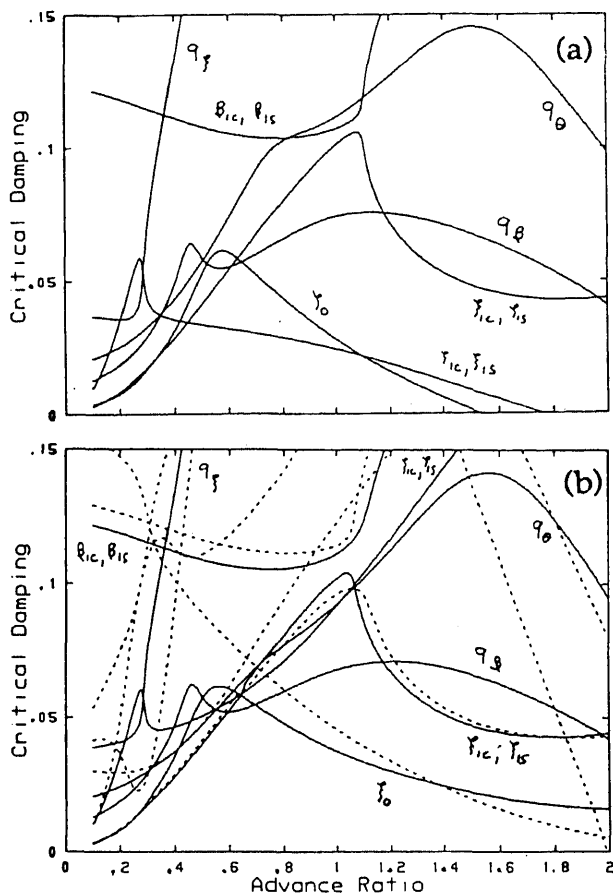
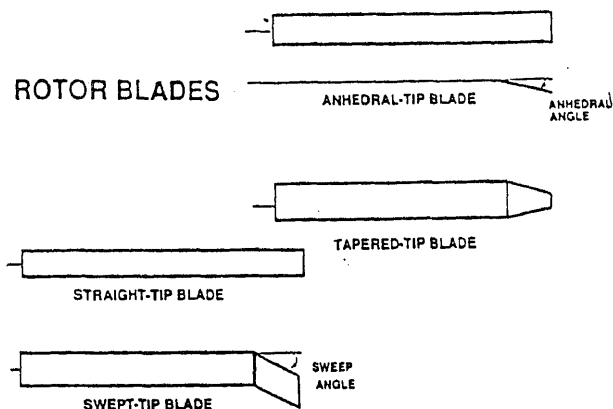


Figure 25. (a) Pylon whirl flutter results (baseline) (Vorwald & Chopra 1991). (b) Influence of controller on whirl flutter stability (Vorwald & Chopra 1991).

## 6. Advanced geometry rotors

To enhance the aerodynamic performance of rotors, in particular at high speed flight, advanced-geometry blades are now frequently used (figure 26). These include such features as specialized airfoils, tapered planform and thickness, swept-tips, anhedral-tips, and nonlinear twist distribution. For example, the incorporation of tip sweep reduces the incident tip Mach number normal to the chord and helps reduce the power requirement by reducing the transonic drag rise on the advancing side of the disk. Advanced tips also appear attractive for reduction of vibration and hub loads, for noise reduction, and for enhancement of aeromechanical and flight stability. However, with these advanced-geometry blades, three-dimensional aerodynamic effects become more important, and also, structural analyses need to be refined. Many of the current rotorcraft codes cannot precisely model these advanced geometry blades.

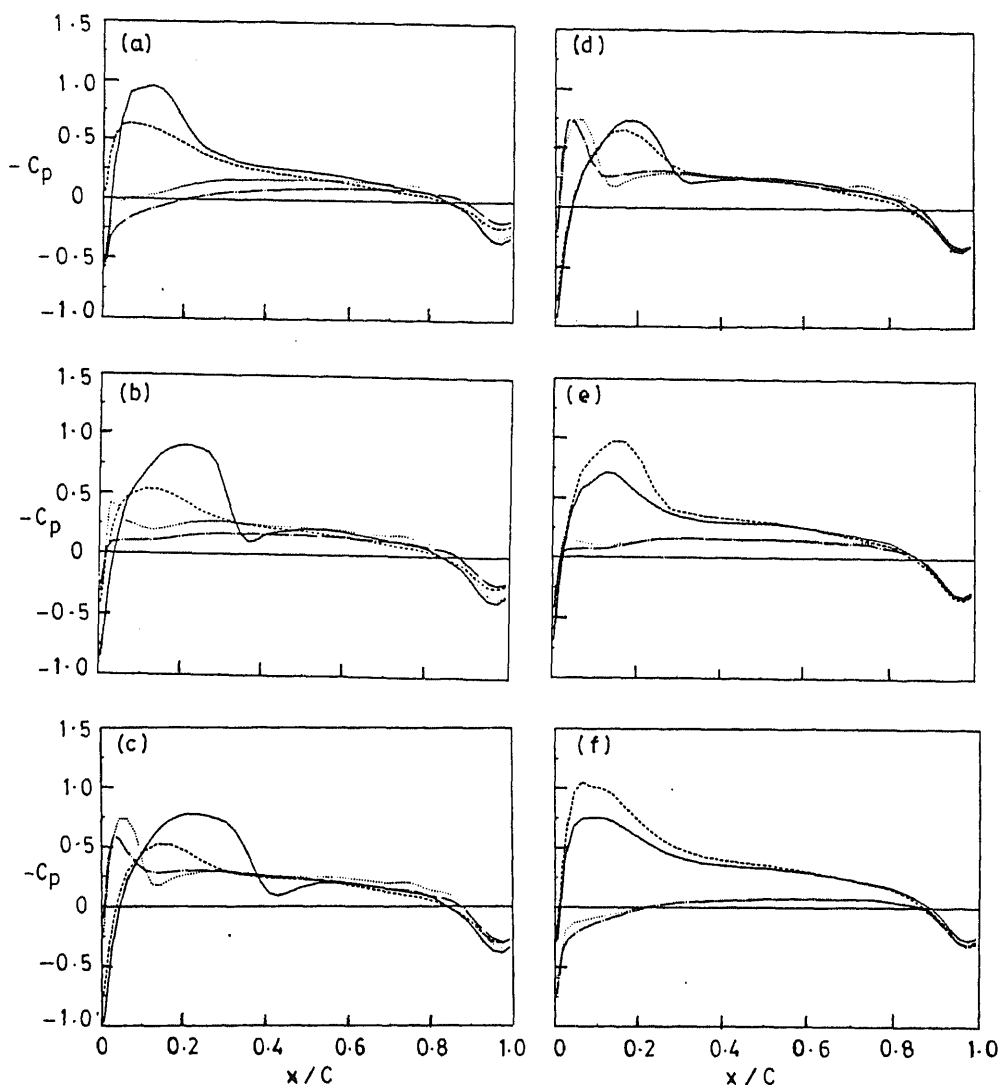
Several researchers involved with rotary-wing aerodynamics have investigated the aerodynamic characteristics of rigid blades with advanced tips (Philippe & Vuillet 1983; Desopper 1985; Desopper *et al* 1989). For example, Desopper 1985, using a detailed CFD method, analysed the flow field and pressure distribution on an isolated



**Figure 26.** Advanced tip-rotors.

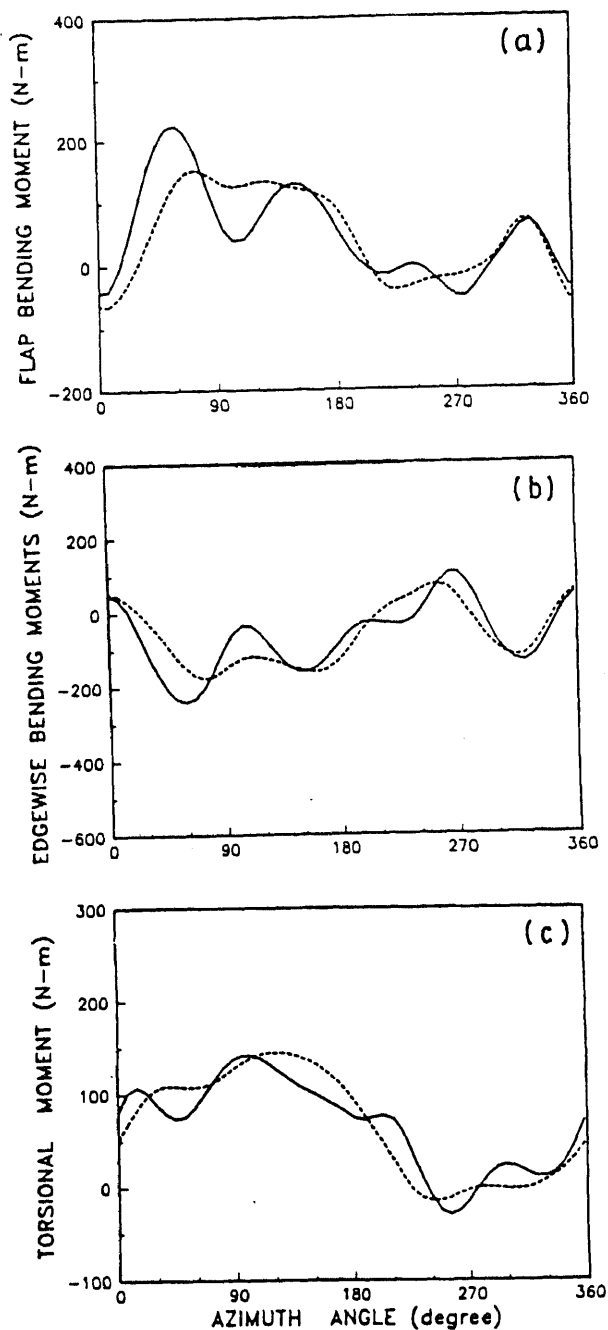
nonlifting blade with advanced tips. It was shown that the use of a combination of sweep and anhedral angle for the tip can reduce the power requirement for the rotor. This study recommended that careful dynamic and structural modelling of the blade be a prerequisite for accurate aerodynamic predictions. Noonan (1985) examined the influence of blade taper ratio and taper initiation point on rotor performance. Using quasisteady aerodynamics, a combination of blade taper, twist and airfoil shape can significantly reduce rotor power.

Some researchers involved with rotor dynamics have addressed the modelling of advanced tips (Panda 1987; Celi & Friedman 1988a; Benquet & Chopra 1989; Kim & Chopra 1992). Celi & Friedmann (1988a) developed a finite element approach to refine the structural representation of swept-tip blades, and analysed the aeroelastic response and stability of the rotor in forward flight. Comparison of results calculated using an equivalent sweep approximation (by shifting chordwise aerodynamic centre and centre of gravity for tip) and a refined structural model showed considerable discrepancy. They pointed out that the routinely-followed equivalent approximation for the tip is incorrect and may lead to erroneous conclusions. It was shown that tip sweep introduces flap-torsion and lag-axial couplings which may lead to aeroelastic instability associated with frequency coalescence. Panda (1987) derived general transformation and constraint relations between two blade elements joined at an angle to each other. Using these relations, Benquet & Chopra (1989) developed a finite element formulation to analyse the response of rotors with swept and anhedral tips. Celi & Friedmann (1988) and Benquet & Chopra (1989) used quasisteady aerodynamics and only linear transformation relations between the tip and blade were included. Kim & Chopra (1992) refined the advanced-tip analysis by including detailed three-dimensional aerodynamics and nonlinear transformation relations. A finite element rotor dynamic analysis was coupled with finite difference aerodynamic analysis, based on transonic small disturbance theory, to calculate the response, structural bending and loads of elastic blades with tip sweep, anhedral and planform taper. It was shown that tip sweep introduces a kinematic axial-lag coupling and a straightening effect of the centrifugal forces that significantly influence the lag dynamics. For a swept-tip blade, the intensity of transonic flows on the advancing side of the rotor is much reduced as compared to a straight-tip blade and this can help improve rotor power requirement and acoustic signatures (figure 27). Tip anhedral has a considerable influence on flap dynamics. The inclusion of three-dimensional aerodynamics in



**Figure 27.** Chordwise pressure distributions of swept-tip ( $30^\circ$ ) and straight blades at  $r/R = 0.98$ ,  $\mu = 0.378$ ,  $C_T/\sigma = 0.064$ ,  $\text{tip} = 0.1R$  (Kim & Chopra 1992). — Lower surface (swept tip); --- upper surface (swept tip); ... lower surface (straight tip); — upper surface (straight tip).  $\Psi = 30^\circ$  (a),  $60^\circ$  (b),  $90^\circ$  (c),  $120^\circ$  (d),  $150^\circ$  (e), and  $180^\circ$  (f).

calculating aerodynamic loads is quite important in order to predict accurately the blade dynamics and loads of an advanced-tip rotor, especially in high-speed flight conditions (figure 28). With a careful selection of tip sweep, anhedral and planform taper, one may design an optimum blade that could improve the performance of a helicopter significantly.



**Figure 28.** Effect of 3-D aerodynamics on structural bending for a swept-tip blade, tip = 0.1R,  $\mu = 0.378$ ,  $C_T/\sigma = 0.064$ . (a) Flap bending, (b) lag bending and (c) torsional moments at 0.8 radius (Kim & Chopra 1992). --- 30° sweep (2-D aero) and — 30° sweep (3-D aero).

## 7. Optimization methodology

Helicopters are susceptible to high vibrations, initiated primarily at the rotor blades. One direct approach to reduce vibration is to design a rotor which inherently produces low oscillatory hub leads. By making an optimum selection of structural,

inertial and aerodynamic characteristics of the blades, it appears possible to minimize the source of vibration and keep the blades aeroelastically stable. An automated methodology used to accomplish this objective is referred to as "aeroelastic optimization."

Aeroelastic optimization has received considerable attention in the fixed-wing field (Ashley 1982). With an enhanced understanding of the dynamics of rotary-wing systems, it is now becoming feasible to apply aeroelastic optimization to rotorcraft. The potential of structural optimization is further expanded with the application of composites in blade construction, which permits great flexibility in tailoring structural characteristics. Also, with the availability of refined optimization algorithms and the substantially increased capability of computers, it is becoming more attractive to implement aeroelastic optimization in the design of complex rotor systems.

Recently, there have been some attempts to apply aeroelastic optimization to helicopter rotors to minimize vibration (for a recent review see Friedman 1991). Blackwell (1983) made a systematic sensitivity study to examine the influence of blade stiffness, spanwise mass distribution, chordwise location of blade centre of gravity, twist, tip sweep and airfoil camber on oscillatory hub loads of a four-bladed articulated rotor. He pointed out the potential of vibration reduction through an optimum blade design. Taylor (1982) considered the vibration reduction of a helicopter through the minimization of a modal shaping parameter. The response of a particular mode to the aerodynamic excitation was minimized by adjusting the mode shape. It is a restrictive approach to reduce vibration. His study was based on simple linear modal analysis, but showed the potential of structural optimization in the rotorcraft field.

Friedmann & Shanthakumaran (1984) applied a modern structural optimization procedure using a sequential unconstrained minimization technique (Miura & Schmit 1979) to reduce oscillatory hub loads for a four-bladed hingeless rotor. They imposed constraints on frequency placements and lag mode stability in hover. Through mass and stiffness distributions of blades, a 15–40% reduction in oscillatory vertical hub shear was achieved for a soft-in-plane hingeless rotor. Peters *et al* (1986) applied an optimization technique (CONMIN, Vanderplaats 1973) to obtain optimum rotating natural frequencies by distributions of mass and stiffness properties of rotor blades. Then, optimum design was pursued for a minimum blade weight with a constraint of its flap inertia. However, a comprehensive aeroelastic analysis was not considered, and stability constraints were not imposed in the optimization process. Davis & Weller (1988) used a constrained optimization program called an automated design synthesis (ADS, Vanderplaats *et al* 1984) to solve four different problems: maximizing bearingless rotor in-plane damping, placing blade natural frequencies, minimizing hub vertical shear and minimizing rotor modal vibration indices. This study was an extension of the work of Taylor (1982) coupled with modern optimization techniques. However, aeroelastic stability constraints were not considered in the optimization process. Davis & Weller (1991) also conducted two experimental programmes to verify the optimization results. In the first programme, the edgewise structural stiffness of a bearingless rotor model was optimized to maximize lag damping. In the second programme, the blade spanwise mass distribution and structural stiffness of an articulated rotor were optimized to minimize rotor vibratory loads in forward flight. The experimental results verified the reliability of the selected optimization criteria and the potential of the modal-based analysis. Yen (1985) showed that a significant reduction in oscillatory hub loads could be achieved by structural optimization of

rotor blades, and stressed the importance of the interaction of rotor structural properties with aerodynamic loads for minimization of hub loads. Simplified rotor dynamic equations were used in the analysis. Correlations of limited flight test and wind tunnel data with theory were also discussed. Recently, Chattopadhyay & Walsh (1989b) applied CONMIN to achieve minimum weight design of a helicopter blade subject to constraints on fundamental coupled flap-lag frequencies and on the auto-rotational inertia. CAMRAD was used to carry out modal analysis of blades. Significant reduction of blade weight was made possible for both rectangular and tapered box beam structures. In this paper, aerodynamic forces were not included. Recently, Chattopadhyay & Walsh (1989a) modified their formulation to include aerodynamic and dynamic coupling in forward flight. Results of this optimization study showed that significant reductions in blade weight and 4/rev vertical hub shear, as well as in oscillatory blade airloads and total rotor power, were possible through the redistribution of the structural characteristics of blades. Celi & Friedmann (1988b) carried out a comprehensive optimization study using CONMIN (Vanderplaats 1973) to minimize oscillatory vertical hub shear for a hingeless rotor with both straight and swept tip blades. They imposed constraints on frequency placements and blade stability in hover. This study showed that the introduction of the swept tip could be beneficial in reducing helicopter vibrations. Celi (1991) applied structural optimization to improve the flight mechanics of a hingeless rotor helicopter. Optimizing the distribution of blade torsional stiffness and cross-sectional offsets (c.g. and elastic axis), the phugoid motion was stabilized while imposing constraints on aeroelastic stability, peak-to-peak bending stresses and longitudinal cyclic response.

Aeroelastic optimization of a system essentially consists of determining the optimum values of design variables that minimize the objective function and satisfy certain aeroelastic and geometric constraints. One of the key elements of an optimization analysis is the calculation of the gradients of the objective function, such as hub loads, structural bending and blade response, and the gradients of aeroelastic constraints such as eigenvalues, with respect to design variables. Rotor dynamics is complex and involves nonlinear inertial, elastic and aerodynamic forces. Most of the existing rotor optimization studies use finite difference methods to calculate sensitivity derivatives, and therefore involve substantial computation time to obtain an optimum solution. Because of prohibitively large computation time, these studies are restrictive in terms of objective functions, dynamic constraints and the number of design variables. Recently, Celi & Friedmann (1988) addressed this issue by building a sequence of approximate optimization problems. This method reduced the total number of function evaluations needed for the calculation of sensitivity derivatives of the objective function and constraints, compared with the conventional finite difference methods.

Lim & Chopra (1990) have made a concerted effort to efficiently calculate sensitivity derivatives of hub loads and stability eigenvalues using a direct analytical approach (chain-rule differentiation). An innovative formulation was developed for the calculation of the sensitivity derivatives and it formed an integral part of regular response and stability solutions. The derivatives were therefore obtained at a fraction of the computation time compared to the frequently adopted finite difference methods (figure 29). For reduction of helicopter vibration, the objective function involved the minimization of one or more components of oscillatory hub loads. In the case in which more than one component was involved with the objective function, suitable weighting functions were incorporated. The constraints involved the aeroelastic

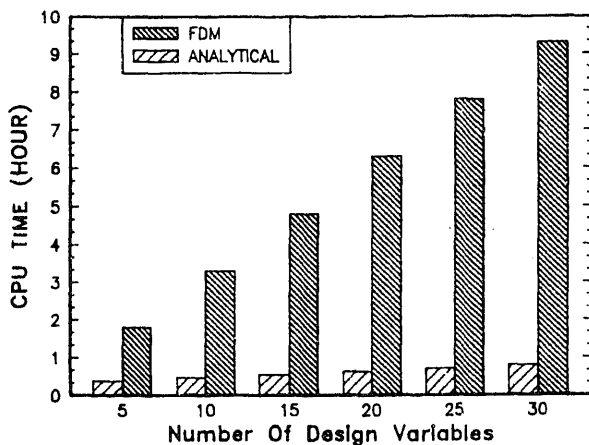


Figure 29. Comparison of CPU time for design sensitivity analysis (Lim & Chopra 1991).

stability of the blade in forward flight, placement of rotating natural frequencies and autorotational performance of the rotor. For the optimization analysis (Lim & Chopra 1989, 1991) an automated optimization process was developed by coupling rotor dynamic analysis and a standard modern optimization code called CONMIN. Optimum solutions were calculated for two types of blade structural representations. One was a generic rotor blade whose structural properties are described in terms of blade stiffness, regardless of the cross-sectional details. In the second type, the blade structural characteristics are defined in terms of a spar geometry of a closed-cell box

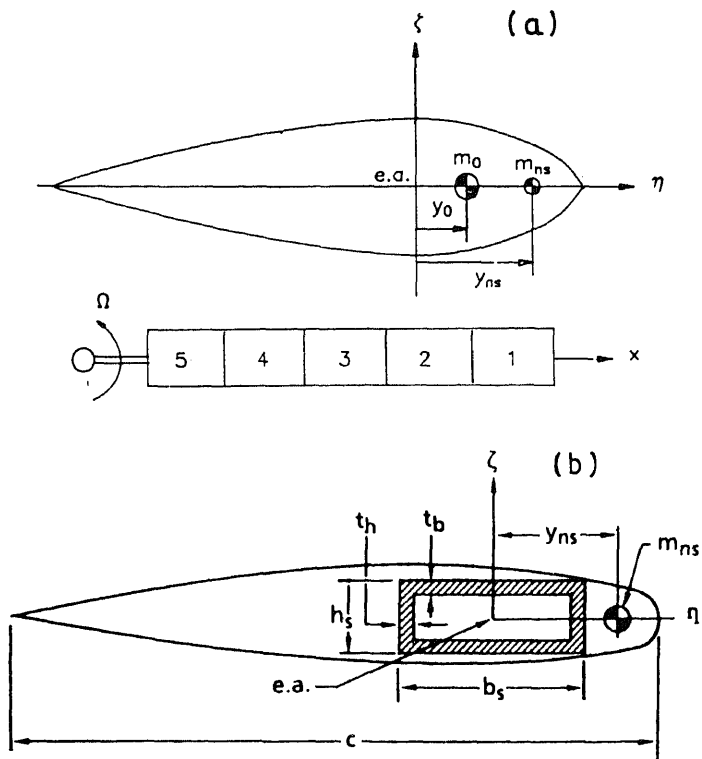


Figure 30. (a) Generic blade and beam elements. (b) Blade cross-section with a box beam.



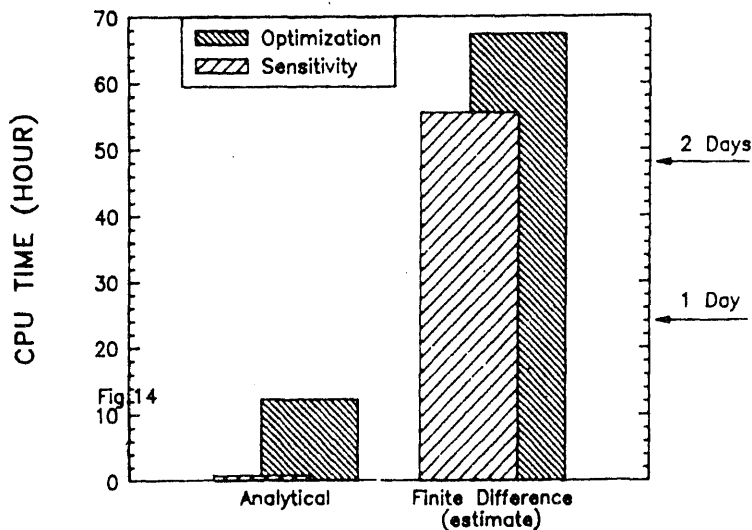


Figure 31. Comparison of CPU time for the optimization procedure (Lim & Chopra 1989).

beam (figure 30). Structural optimization was carried out on a four-bladed, soft-in-plane hingeless rotor to minimize six components of oscillatory hub loads while maintaining lag mode stability in forward flight and certain behaviour constraints. It was shown that the direct analytical approach for sensitivity derivatives resulted in an 80% reduction in total CPU time compared with commonly used finite difference approaches (figure 31). The minimization of 4/rev vertical hub shear alone is not sufficient to reduce vibration. The best design solution is achieved by distribution of nonstructural masses and blade-bending stiffnesses (flap, lag and torsion), and removing the blade c.g. as a design variable (figure 32). Optimum solutions resulted in reductions of 25–77% for a generic blade, and 30–50% for the box-beam blade relative to baseline values of the objective function that was composed of all six components of hub

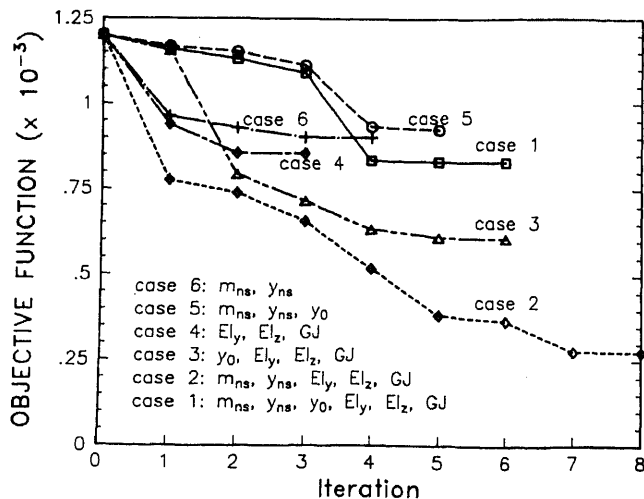
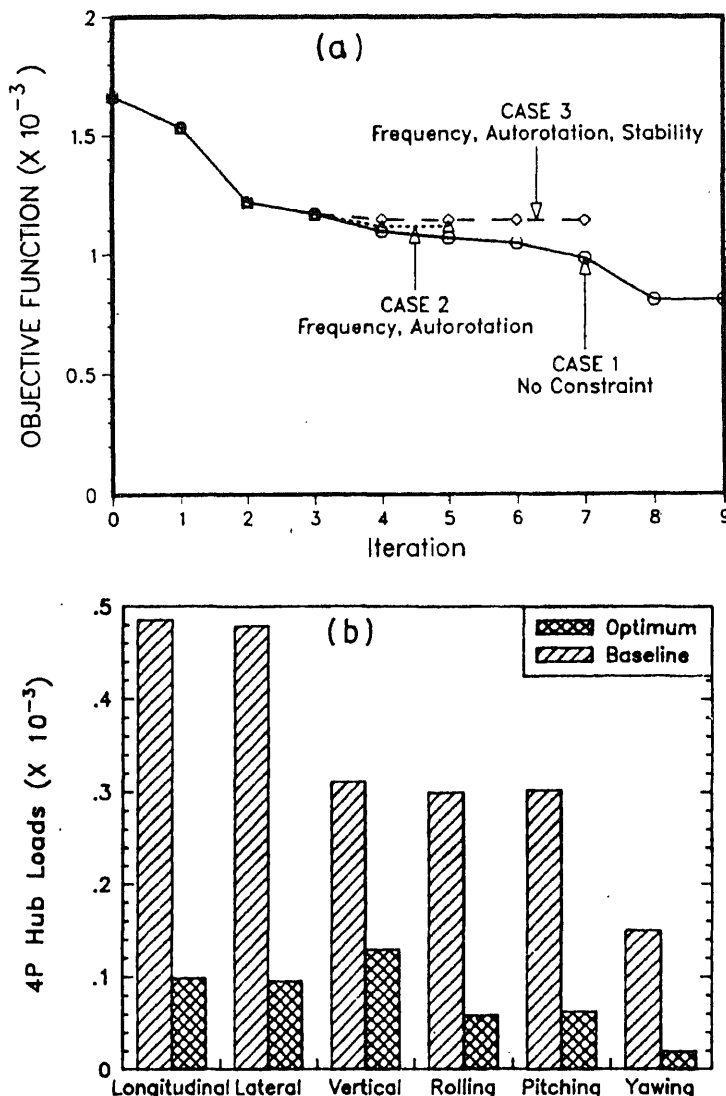


Figure 32. Optimization iteration history of objective function for a generic blade,  $C_T/\sigma = 0.07$ ,  $\mu = 0.3$  (Lim & Chopra 1989).



**Figure 33.** (a) Optimization iteration history of objective function for a box beam blade,  $C_T/\sigma = 0.07$ ,  $\mu = 0.3$ , soft-in-plane hingeless rotor (Lim & Chopra 1991). (b) Optimum 4/rev hub forces and moments (Lim & Chopra 1991).

loads (figure 33). Ganguli & Chopra (1992a) showed that using advanced geometry blades such as including distributed sweep, anhedral, pretwist and planform taper, it is possible to achieve additional reduction of 10–15% of oscillatory hub loads in an optimum rotor design. Sensitivity and optimization analysis of a composite rotor was carried out for a box-beam laminated spar (Ganguli & Chopra 1992b). Starting from an initially infeasible design, the optimum design solution for a four-bladed soft-in-plane rotor with structurally tailored pitch-lag coupling shows an increase on lag damping of over 100%.

Utilizing the structural optimization methodology, there is great potential for

designing rotor blades to enhance the efficiency of rotorcraft. It may further expand the optimization potential through the application of composite tailoring and advanced geometry tips. It will be important to check these gains by testing dynamically-scaled optimized rotor models at different flight conditions.

## 8. Smart structures technology

As stated previously, helicopters are susceptible to high vibratory loads, aeromechanical instabilities, excessive noise levels, poor flight stability characteristics, and high dynamic stresses. To reduce these problems to an acceptable level, numerous passive and active devices, and many *ad hoc* design fixes, are resorted to with resultant weight penalties and reduced payloads. The primary source for all these problems is the nonsteady and complex aerodynamic environment in which the rotor must operate. To counter some of these deficiencies, and also to further expand the flight capabilities of military and civilian helicopters, many new design modifications and devices are being contemplated. These appear to show incremental and modest gains in terms of performance improvement and reduction in operating costs. If the objective is to achieve 'a jet smooth ride' with helicopters at a comparable operating cost, for example, one has to try revolutionary ideas. One innovative idea that may give a substantial jump in performance at a small price is to apply the technology of smart structures to rotorcraft. For such an application, numerous light-weight sensors and actuators are embedded or surface-mounted at different stations on the blades, transmission shafts and airframe, and optimal distributed forces applied with the help of modern control theory. At this stage, the technology of smart structures is primitive and requires a focused basic research effort that will help clarify the projected gains. It is therefore necessary to separate reality from fantasy, practicality from hypothesis, and full-scale possibilities from laboratory models.

A smart structure involves distributed actuators and sensors, and one or more microprocessors that analyse the responses from the sensors and use distributed-parameter control theory to command the actuators to apply localized strains. A smart structure has the capability to respond to a changing external environment (such as loads and shape change) as well as to a changing internal environment (such as damage or failure). Many types of actuators and sensors are being considered, such as piezoelectric materials, shape-memory alloys, electrostrictive materials, magnetostrictive materials, electro-rheological fluids and fibre optics. These can be integrated with main load-carrying structures by surface bonding or embedding without causing any significant changes in the structural stiffness of the system. Among these, piezoelectrics are the most popular. They undergo surface elongation (strain) when an electric field is applied across them and produce voltage when surface strain is applied, and thus can be used both as actuators and sensors. Among other materials, shape-memory alloys are gaining rapid recognition as actuators because of the possibility of achieving large excitation forces, and fibre optics are becoming popular as sensors because they can be easily embedded in composite structures with little effect on the material integrity.

Recently, there has been an increase in smart structures research activities. Much of this work is focused on the application of piezoelectric technology to space-related systems, such as the control of vibration of large space structures (Crawley & de Luis 1987; Anders & Rogers 1991) and for stable bases for precision pointing in space

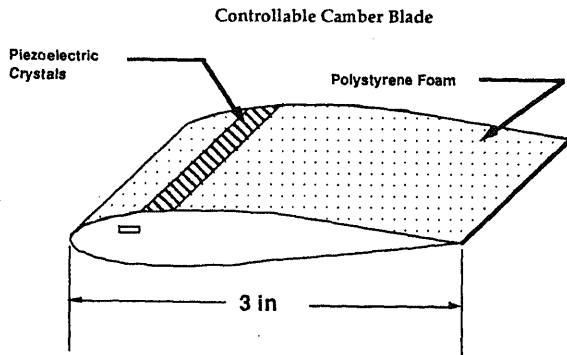


Figure 34. Intelligent rotor model.

(telescope, mirrors etc., Wada *et al* 1990). Also, there are a few preliminary applications to the field of fixed-wing aircraft, such as controlling wing twist and camber for flutter suppression (Scott & Weisshaar 1991) and for controlling structure-borne noise (Atluri & Amos 1989). To date, very little work has been conducted on the use of smart structures for rotorcraft applications. Some of the initial work in applying this emerging technology to the rotor was initiated at Maryland by Barrett (1990). A Froude-scaled two-bladed intelligent rotor model of six-foot diameter was built (figure 34) and tested in the vacuum chamber under rotating conditions. To actively and independently manipulate bending and twist distributions of the blades, directionally attached piezoelectric (DAP) crystals were embedded. Testing of the rotor demonstrated an active tip twist up to  $2^\circ$  at frequencies as high as 150 Hz (about

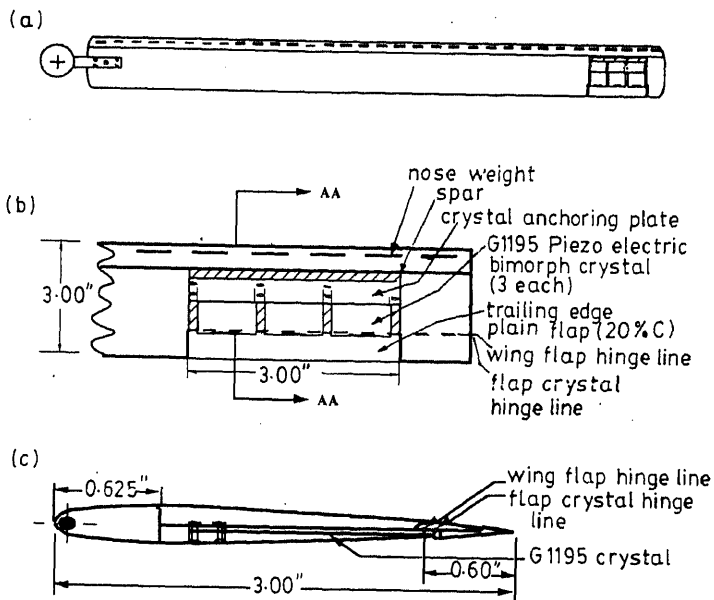
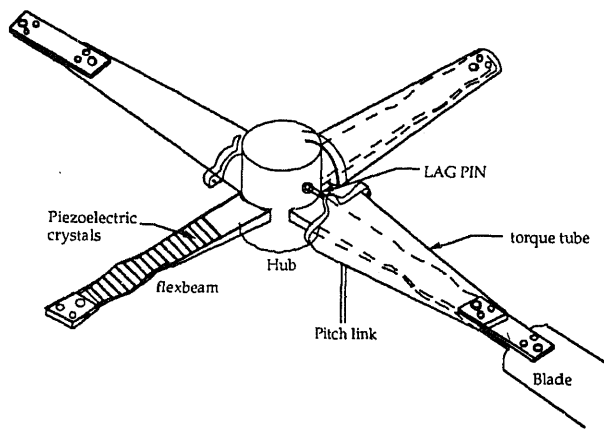


Figure 35. Froude scale rotor model using smart structures technology (Chopra & Samak 1991). (a) Main rotor blade with flap actuator. (b) Details of flap actuator arrangement. (c) Actuator cross-section AA.

10 per rev). Using a simple feedback system, it was shown that the forced flapwise vibration of this rotating blade with tip amplitude of about 10% of radius could be significantly reduced (up to 70%). To improve the magnitude of the actuation forces, an alternate model design with a trailing-edge flap is being pursued (figure 35). Through a specially designed mounting arrangement for the flap, small actuation strains of piezo bimorphs are amplified into angular deflections of the flap. It is possible to further improve the actuation by incorporating beneficial composite couplings in design. Also, a Froude scale bearingless rotor model is being built incorporating the smart structures technology to augment aeromechanical stability (figure 36). At the Massachusetts Institute of Technology, Spangler & Hall (1990) also built a blade model using piezoelectric actuators on a trailing edge flap. Though it was not dynamically scaled or meant to be tested in a rotating environment, it showed the promise of piezo-actuators. At Georgia Tech, Hanagud & Nagesh Babu (1991) also have been trying to apply piezoelectric technology to reduce the forced vibration of a cantilevered beam specimen. Again, it was not a scaled model, nor were issues related to the rotating environment addressed. Unless testing is carried on with a simulated scaled model, one may not be able to appreciate the actuation force requirement for an actual rotor system. Also, there are several unique and complex issues related to the implementation of smart structure technology in rotorcraft, such as feeding high voltage signals to different actuators in the rotating environment, the need for high force-high strain actuators, the need for small-size and light-weight smart structures, and for tight control on the structural integrity of blades. Unless these issues are addressed in research, one may arrive at false conclusions. To carry out a meaningful study on this topic, one must demonstrate this technology on a dynamically scaled rotor model in the rotating environment.

The structural, mechanical and aerodynamic complexity and the multidisciplinary nature of rotorcraft offer many opportunities for the application of smart structures technology with the potential for very substantial payoffs in system effectiveness. The rotor is the key subsystem, setting the current limits on vehicle performance, handling qualities and reliability. Since the rotor is also a flexible structure, changes in shape, mechanical properties and stress/strain fields can be imposed upon it. These in turn will alter the vibratory modes, aeroelastic interactions, aerodynamic properties, and dynamic stresses of the rotor and fuselage. Smart structures technology will enable these imposed changes to be tailored to conditions sensed in the rotor itself.



**Figure 36.** Aeromechanical stability augmentation in a bearingless rotor (6 ft dia) using smart structures technology.

Furthermore, because the smart actuators and sensors can be distributed over each individual rotor blade, control can be imposed over a much larger bandwidth than with current swashplate-based controls which are limited to  $N/\text{rev}$  for an  $N$ -bladed rotor. This opens up a hitherto unavailable domain for vibrational control, aero-mechanical stability augmentation, handling quality enhancement, and reduction of acoustic signatures. The use of smart structures also offers the prospect of sensing structural damage in the rotor structure and in other critical components. The pilot can then be alerted, enabling him/her to take load alleviation action. There also exists the potential for altering the stress field following damage, using smart materials. This could provide a degree of self-repair. A further very promising application of smart structures is to control the critical frequency of drive shafts.

Basic smart structures technology also needs to be advanced if the potential for use on rotocraft is to be fully realized. Most critically, we need new concepts for smart actuators with high strain rates, large bandwidths and minimum impact on the host structure. We need better understanding of the transfer functions and failure modes of the smart devices themselves. We need new, rapid, robust algorithms for control systems involving multiple, distributed sensors and actuators.

In order to realise the potential advantages of smart structures for rotorcraft, an integrated research effort is required. This research should investigate the exploitation of smart structure concepts in a realistic environment, emulating that of the rotorcraft, and advance the underlying generic smart structure concepts and technologies that could benefit rotorcraft. This integrated approach is essential if the problems limiting the application of smart structures are to be identified, and if fundamental research efforts are to be properly focused.

## 9. Conclusions and recommendations for future work

*Hingeless and bearingless rotors:* For an accurate analysis of bearingless rotors, one needs to incorporate finite element discretization, and modelling of redundant load paths and large elastic deflections. These rotors suffer from air resonance instability, and to stabilize this, innovative designs must be explored to achieve beneficial aeroelastic couplings.

*Composite rotors:* For analysis of composite blades, nonclassical effects such as section warping, transverse shear, and in-plane elasticity become important and must be included. For open-section beams, modelling of warping constraints must also be incorporated. Systematic experimental testing of simple composite beams in rotating and nonrotating environments is needed to validate analyses. The potential for tailoring of structural couplings to minimize blade stresses and vibration and to increase aeromechanical stability must be explored.

*Circulation control rotor:* Trailing-edge blowing has a powerful influence on the dynamics and flight mechanics of fixed-wing and rotary-wing airplanes. For proper CC rotor analysis, the perturbation, unsteady and pneumodynamics effects associated with blowing must be incorporated. Circulation control aerodynamics causes some new aeroelastic instabilities, such as blowing flutter and control reversal, and these must be explored systematically. The state of art on CC technology is quite primitive

and reliable mathematical tools need to be developed before it can be exploited in rotorcraft.

**Tilt-rotors:** The dynamics of a tilt-rotor is more involved than a conventional rotor because of complex couplings due to wing and pylon motion, high inflow ratio in airplane mode, highly twisted blades and tilting of main rotors. For an accurate prediction of the dynamic instability, it is important to include blade bending flexibility, proper modelling of rotor hub, pylon and wing motion, and proper representation of gimbal and rotor-speed degrees of freedom. The potential of an active feedback system to augment whirl flutter instability must be investigated.

**Advanced geometry rotors:** For accurate analysis of advanced geometry rotors, it is necessary to develop special finite elements for the tip including nonlinear transformation relations between main blade and tip and three dimensional aerodynamics. Advance tips such as swept, anhedral and tapered-planform have significant influence on dynamics, performance and acoustic signatures. Utilizing automated optimization methodology, tip geometry must be explored to improve rotor efficiency.

**Structural optimization methodology:** For practical applications of structural optimization in rotorcraft, it is essential to develop an efficient sensitivity analysis. For minimization of rotorcraft vibration, it is needed to minimize an objective function consisting of all rotor hub loads. The domain of optimization must be expanded to cover flight mechanics and basic performance and to utilize composite tailoring and advanced geometry tips.

**Smart structures technology:** At this time, this technology is primitive and a focussed research effort is needed before it can be exploited in rotorcraft. It is necessary to develop light-weight actuators with high strain and band width, formulate robust control algorithms for distributed actuators and sensors and check implementation in rotating environments.

## References

- Anders W S, Rogers C A 1991 Design of a shape memory alloy deployment hinge for reflector facets. *Proceedings of the AIAA/ASME/ASCE/AHS/ASC 32nd Structures, Structural Dynamics and Materials Conference*, Baltimore, Maryland
- Ashley H 1982 On making things the best – aeronautical uses of optimization. *J. Aircraft* 19: 5–28
- Atluri S N, Amos A K (eds) 1989 *Large space structures: Dynamics and control* (Berlin: Springer Verlag)
- Barnes D R, Bill F A, Wilkerson J B 1979 Circulation control flight demonstrator test program. *Proceedings of the 35th Annual Forum of the American Helicopter Society*, Washington, DC
- Barrett R 1990 Intelligent rotor blade and structures development using directionally attached piezoelectric crystals. Master of Science thesis, Department of Aerospace Engineering, University of Maryland, College Park
- Bauchau O A, Hong C H 1988 Nonlinear composite beam theory. *J. Appl. Mech.* 55: 156–163
- Benquet P, Chopra I 1989 Calculated dynamic response and loads for an advanced tip rotor in forward flight. *Proceedings of the Fifteenth European Rotorcraft Forum*, Amsterdam, Netherlands

- Bir G S, Chopra I 1991 Effect of nonlinear elastomeric dampers on aeromechanical stability of helicopters. *Proceedings of the International Technical Specialists Meeting on Rotorcraft Basic Research*, Georgia Tech, Atlanta
- Blackwell R H 1983 Blade design for reduced helicopter vibration. *J. Am. Helicopter Soc.* 28: 33–41
- Celi R 1991 Optimization aeroelastic design of helicopter rotors for longitudinal handling qualities improvement. *J. Aircraft* 28: 49–57
- Celi R, Friedmann P P 1988a Aeroelastic modeling of swept tip rotor blades using finite elements. *J. Am. Helicopter Soc.* 33: 43–52
- Celi R, Friedmann P P 1988b Structural optimization with aeroelastic constraints of rotor blades with straight and swept tips. *Proceedings of the 29th AIAA Structures, Structural Dynamics and Materials Conference*, Williamsburg, Virginia
- Chandra R, Chopra I 1991a Experimental and theoretical analysis of composite I-beams with elastic couplings. *AIAA J.* 29: 2197–2205
- Chandra R, Chopra I 1991b Vibration characteristics of composite I-beams with elastic couplings under rotation. *Proceedings of 47th Annual Forum of American Helicopter Society*, Phoenix, Arizona
- Chandra R, Chopra I 1992a Influence of elastic couplings on vibration characteristics of thin-walled composite box beams under rotation. *J. Aircraft* 29: 657–664
- Chandra R, Chopra I 1992b Structural response of composite beams and blades with elastic couplings. *Compos. Eng., Int. J.* 2: 347–374
- Chandra R, Stemple A D, Chopra I 1990 Thin-walled composite beams under bending, torsional and extensional loads. *J. Aircraft* 27: 619–626
- Chattopadhyay A, Walsh J L 1989a Integrated aerodynamic/dynamic optimization of helicopter rotor blades. *Proceedings of the 30th AIAA Structures, Structural Dynamics and Materials Conference*, Mobile, Alabama
- Chattopadhyay A, Walsh J L 1989b Minimum weight design of helicopter rotor blades with frequency constraints. *J. Am. Helicopter Soc.* 34: 77–82
- Cheeseman I C 1967 The application of circulation control by blowing to helicopters. *J. R. Aeronaut. Soc.* 71: July
- Chen W Y, Chopra I 1991 Aeromechanical stability of helicopters in forward flight. *Proceedings of the International Technical Specialists Meeting on Rotorcraft Basic Research*, Georgia Tech, Atlanta
- Chopra I 1984 Aeroelastic stability of an elastic circulation control rotor blade in hover. *Vertica* 8: 353–371
- Chopra I 1985 Aeroelastic stability of a bearingless circulation control rotor blade in hover. *J. Am. Helicopter Soc.* 30: 40–47
- Chopra I 1988 Aeroelastic stability of a bearingless circulation control rotor in forward flight. *J. Am. Helicopter Soc.* 33: 60–67
- Chopra I 1990 Perspectives in aeromechanical stability of helicopter rotors. *Vertica* 14: 457–508
- Chopra I, Johnson W 1979 Flap-lag-torsion aeroelastic stability of circulation control rotors in hover. *J. Am. Helicopter Soc.* 23: 37–46
- Chopra I, Samak D K 1991 Development of an intelligent rotor. *Conference on Active Materials and Adaptive Structures*, Arlington, Virginia
- Crawley E F, de Luis J 1987 Use of piezoelectric actuators as elements of intelligent structures. *AIAA J.* 25: 1373–1385
- Davis M W, Weller W M 1988 Application of design optimization techniques to rotor dynamics problems. *J. Am. Helicopter Soc.* 33: 42–50
- Davis M W, Weller W H 1991 Helicopter rotor dynamics optimization with experimental verification. *J. Aircraft* 28: 38–48
- Desopper A 1985 Study of unsteady transonic flow on rotor blade with different tip shapes. *Vertica* 9: No. 3, July
- Desopper A, Lafon P, Ceroni P, Philippe J J 1989 Ten years of rotor flow studies at ONERA. *J. Am. Helicopter Soc.* 34: 41–51
- Dull A L, Chopra I 1988 Aeroelastic stability of bearingless rotors in forward flight. *J. Am. Helicopter Soc.* 33: 38–46
- Englar R J, Applegate C A 1984 Circulation control – A bibliography of David Taylor Naval



- Ship Research and Development Center Research and selected outside references. Report DTNSRDC 84/052, AD A146-966, September
- Englar R F, Trobaugh L A, Hemmerly R A 1978 STOL potential of the circulation control wing for high performance aircraft. *J. Aircraft* 15: 175-181
- Friedmann P P 1991 Helicopter vibration reduction using structural optimization with aeroelastic/multidisciplinary constraints - A survey. *J. Aircraft* 28: 8-21
- Friedmann P P, Shanthakumaran P 1984 Optimum design of rotor blades for vibration reduction in forward flight. *J. Am. Helicopter Soc.* 29: 70-80
- Ganguli R, Chopra I 1992a Aeroelastic optimization of an advanced geometry helicopter rotor. *Proceedings of the 33rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Dallas, Texas
- Ganguli R, Chopra I 1992b Aeroelastic optimization of a composite helicopter rotor. *Fourth AIAA/USAF/NASA/OAI symposium on Multidisciplinary Analysis and Optimization Conference*, Cleveland, Ohio
- Ghee T A, Leishman J G 1990 Effects of unsteady blowing on the lift of a circulation controlled cylinder. *J. Am. Helicopter Soc.* 35: 22-31
- Haas D J, Chopra I 1988 Static aeroelastic characteristics of circulation control wings. *J. Aircraft* 25: 948-954
- Haas D J, Chopra I 1989a Flutter of circulation control wings. *J. Aircraft* 26: 373-381
- Haas D J, Chopra I 1989b Aeroelastic stability of aircraft with circulation control wings. *Proceedings of the 30th AIAA Structures, Structural Dynamics and Materials Conference*, Mobile, Alabama
- Hanagud S, Nagesh Babu G L 1991 Smart structures in the control of airframe vibrations. *Proceedings of the International Technical Specialists' Meeting on Rotorcraft Basic Research*, Georgia Institute of Technology, Atlanta, Georgia
- Hodges D H, Atilgan A R, Fulton M V, Rehfield L W 1989 Dynamic characteristics of composite beam structures. *Proceedings of the National Technical Specialists Meeting on Rotorcraft Dynamics* (Arlington, TX: Am. Helicopter Soc.)
- Hong C H, Chopra I 1985 Aeroelastic stability of a composite rotor blade. *J. Am. Helicopter Soc.* 30: 57-67
- Hong C H, Chopra I 1986 Aeroelastic stability analysis of a composite bearingless rotor blade. *J. Am. Helicopter Soc.* 31: 29-35
- Jang J, Chopra I 1988 Air resonance of an advanced bearingless rotor in forward flight. *Second International Conference on Rotorcraft Basic Research*, College Park, Maryland
- Jang J, Chopra I 1989 Ground and air resonance of an advanced bearingless rotor in hover. *J. Am. Helicopter Soc.* 33: 20-29
- Johnson W 1974a Dynamics of tilting proprotor aircraft in cruise flight. NASA TN D-7677
- Johnson W 1974b Analytical model for tilting proprotor aircraft dynamics, including blade torsion and coupled bending modes, and conversion mode operation. NASA TM X-62,369
- Johnson W 1975 Analytical modeling requirements for tilting proprotor aircraft dynamics. NASA TN D-8083
- Johnson W 1980 A comprehensive analytical model for rotorcraft aerodynamics and dynamics. NASA TM-81182
- Johnson W 1981 Development of a comprehensive analysis for rotorcraft. *Vertica* 5: 99-130
- Johnson W 1984 An assessment of the capability to calculate tilting proprotor aircraft performance, loads, and stability. NASA Technical Paper 2291
- Johnson W 1985 Recent developments in the dynamics of advanced rotor systems. NASA Technical Memorandum 86669
- Johnson W, Chopra I 1979 Calculated hovering helicopter flight dynamics with a circulation-controlled rotor. *J. Aircraft* 16: 124-128
- Kim K C, Chopra I 1992 Aeroelastic analysis of rotor blades with advanced tip shapes. *J. Am. Helicopter Soc.* 38: 15-30
- Kosmatka J B, Friedmann P P 1988 Structural dynamic modeling of advanced composite propellers by the finite element method. *Proceedings of the 28th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Mobile, Alabama
- Kvaternik R G 1973 *Studies in tilt-rotor VTOL aircraft aeroelasticity*. Ph D dissertation, Case Western University
- Lim J W, Chopra I 1989 Aeroelastic optimization of a helicopter rotor. *J. Am. Helicopter Soc.* 34: 52-62

- Lim J W, Chopra I 1990a Response and hub loads sensitivity analysis of a helicopter blade. *AIAA J.* 28: 75–82
- Lim J W, Chopra I 1990b Stability sensitivity analysis for aeroelastic optimization of a helicopter rotor. *AIAA J.* 28: 1089–1097
- Lim J W, Chopra I 1991 Aeroelastic optimization of a helicopter rotor using efficient sensitivity analysis. *J. Aircraft* 28: 29–37
- Linden A W, Biggers J C 1985 X-Wing potential for navy applications. *Proceedings of the 41st Annual Forum of the American Helicopter Society*, Fort Worth, Texas
- Logan A H 1982 Design and flight test of the no tail rotor (NOTAR) aircraft. *Proceedings of the 38th Annual Forum of the American Helicopter Society*, Anaheim, California
- Lorber P F, Carson R G 1989 The aerodynamics of an oscillating jet flap. *J. Am. Helicopter Soc.* 34: 24–32
- McNulty M J, Bousman W G (eds) 1983 *Proceedings of the Integrated Technology Rotor Methodology Assessment Workshop*, sponsored by NASA Ames Research Center and the US Army, NASA Conference Publication 10007, June 21–22
- Minguet P, Dugundji J 1990a Experiments and analysis for composite blades under large deflections. Part 1: Static behavior. *AIAA J.* 28: 1573–1579
- Minguet P, Dugundji J 1990b Experiments and analysis for composite blades under large deflections. Part 2: Dynamic behavior. *AIAA J.* 28: 1580–1588
- Miura H, Schmit L A 1979 NEWSUMT – A Fortran program for inequality constrained function minimization. User's Guide, NASA CR 159070
- Nixon M W 1989 Extensional-twist coupling of composite circular tubes with application to tilt rotor blade design. *Proceedings of the 28th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Mobile, Alabama
- Noonan K W 1985 *Aerodynamic design of a helicopter main rotor blade with consideration of flap-lag flutter in hover*. M S thesis, University of Maryland
- Panda B 1987 Assembly of moderate-rotation finite elements used in helicopter rotor dynamics. Technical Note. *J. Am. Helicopter Soc.* 32: 63–69
- Panda B, Chopra I 1987 Dynamics of composite rotor blades in forward flight. *Vertica* 11: 187–210
- Peters D A, Rossow M P, Korn A, Ko T 1986 Design of helicopter rotor blades for optimum dynamic characteristics. *Comput. Math. Appl.* A12: 85–110
- Philippe J J, Vuillet A 1983 Aerodynamic design of advanced rotors with new tip shapes. *Proceedings of the 39th Annual Forum of the American Helicopter Society*, St. Louis, Missouri, pp. 267–277
- Popelka D, Sheffler M, Bilger J 1985 Correlations of stability test results and analysis for the 1/5 scale V-22 aeroelastic model. *Proceedings of 41st Annual Forum of American Helicopter Society*, Ft. Worth, Texas
- Raghavan V, Pai S, Chopra I 1988 Circulation control airfoils in unsteady flow. *J. Am. Helicopter Soc.* 33: 28–37
- Reader K R, Kirkpatrick D G, Williams R M 1978 Status report on advanced development program utilizing circulation control rotor technology. *Proceedings of the Fourth European Rotorcraft and Powered Lift Aircraft Forum*, Stresa, Italy
- Rehfield L W, Atilgan A R, Hodges D H 1990 Nonclassical behavior of thin-walled composite beams with closed cross-sections. *J. Am. Helicopter Soc.* 35: 42–50
- Scott R C, Weisshaar T A 1991 Controlling panel flutter using adaptive materials. *Proceedings of the AIAA/ASME/ASCE/AHS/ASC 32nd Structures, Structural Dynamics and Materials Conference*, Baltimore, Maryland
- Sivaneri N T, Chopra I 1984 Finite element analysis for bearingless rotor blade aeroelasticity. *J. Am. Helicopter Soc.* 29: 42–51
- Smith E C, Chopra I 1991 Formulation and evaluation of an analytical model for composite box beams. *J. Am. Helicopter Soc.* 36: 23–35
- Smith E C, Chopra I 1992a Aeroelastic response and blade loads of a composite rotor in forward flight. *Proceedings of the 33rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Dallas, Texas
- Smith E C, Chopra I 1992b Aeromechanical stability of helicopters with composite rotors in forward flight. *Proceedings of 48th Annual Forum of American Helicopter Society*, Washington DC

- Spangler R, Hall S 1990 Piezoelectric actuators for helicopter rotor control. *Proceedings of AIAA/ASME/ASCE/AHS/ASC 31st Structures, Structural Dynamics and Materials Conference*, Long Beach, California
- Stemple A D, Lee S W, 1988 A finite element model for composite beams with arbitrary cross sectional warping. *AIAA J.* 26: 1520–1520
- Taylor R B 1982 Helicopter vibration reduction by rotor blade model shaping. *Proceedings of the 38th Annual Forum of the American Helicopter Society*, Anaheim, California
- Vanderplaats G N 1973 CONMIN – A Fortran program for constrained function minimization. User's Manual, NASA TMX 62282
- Vanderplaats G N, Sugimoto H, Sprague C M 1984 ADS-1: A new general purpose optimization program. *AIAA J.* 22: 1458–1460
- Vorwald J G, Chopra I 1991 Stabilizing pylon-whirl flutter on a tilt-rotor aircraft. *Proceedings of AIAA/ASME/AHS/ASC 32nd Structures, Structural Dynamics and Materials Conference*, Baltimore, Maryland
- Wada B K, Fanson J L, Crawley E F 1990 Adaptive structures. *Proceedings of the ASME Winter Annual Meeting*, December
- Wang J M, Chopra I 1990 Bearingless rotor aeromechanical stability measurements and correlations using nonlinear aerodynamics. *16th European Rotorcraft Forum*, Glasgow, Scotland
- Wang J, Chopra I, Samak D K, Green M, Graham T 1989 Theoretical and experimental investigation of aeroelastic stability of an advanced bearingless rotor in hover and forward flight. *Proceedings of the National Technical Specialists' Meeting on Rotorcraft Dynamics*, (Arlington, TX: Am. Helicopter Soc.)
- Wang J, Jang J, Chopra I 1990 Air resonance of hingeless rotors in forward flight. *Vertica* 14: 123–136
- Watkins C B, Reader K R, Dutta S K 1985 Pneumodynamic characteristics of a circulation control rotor model. *J. Am. Helicopter Soc.* 30: 23–31
- Yen J G 1985 Coupled aeroelastic hub loads reductions. AHS/NAI International Seminar, Nanjing, China

## Computational aircraft dynamics and loads

K APPA<sup>1</sup> and J ARGYRIS<sup>2</sup>

<sup>1</sup>Northrop Corporation, Aircraft Division, MS 3969/63, One Northrop Avenue, Hawthorne, California 90250-3277, USA

<sup>2</sup>Institute for Computer Applications, Pfaffenwaldring 27 D-70569 Stuttgart, University of Stuttgart, Germany

**Abstract.** An alternative approach that has the potential to advance classical methods of flight load prediction by combining computational fluid dynamics (CFD), structural flexibility and the interaction of flight control system (FCS) in a multidisciplinary analysis package is described. The method employs the concept of system identification to characterize aircraft dynamics in the state space coordinate system and includes an adaptive control law design methodology. An extended account of the theoretical basis for the new multidisciplinary flight manoeuvre analysis will be presented in one of a seven-volume series on computational mechanics by Argyris and his associates to be published shortly. However, as a precursor to the complete work, a brief account of the theoretical development leading to this loads prediction methodology is included in this paper.

**Keywords.** Flight loads; manoeuvre loads; multidisciplinary computational methods; external loads.

### 1. Introduction

The structural design of aircraft requires the knowledge of external loads acting on individual components. The computation of such loads depends on the flight environment to which a particular aircraft is subjected. For example, commercial (or transport) aircraft is expected to withstand loads due to level flight, gust encounter, landing and take off, and ground handling. On the other hand, a combat aircraft is required to sustain additional loads due to rapid manoeuvres which are several times higher than normal operational loads. The development of load spectra and load envelope depends on a particular aircraft's mission and is typically specified by the procuring agencies.

The critical design flight conditions used to compute load exceedance envelopes such as shear vs. bending moment, and torsion vs. bending moment can be simulated by moving the control sticks at desired rates (see Appa 1991). The modern aircraft design methodology requires that the flight control systems (FCS) be included in

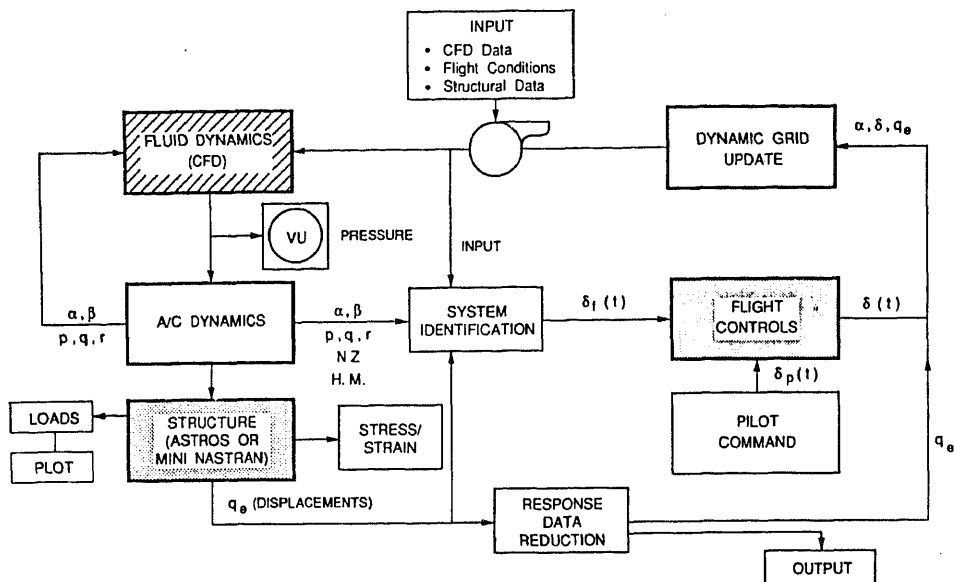


Figure 1. A schematic diagram of computational aircraft dynamics and loads.

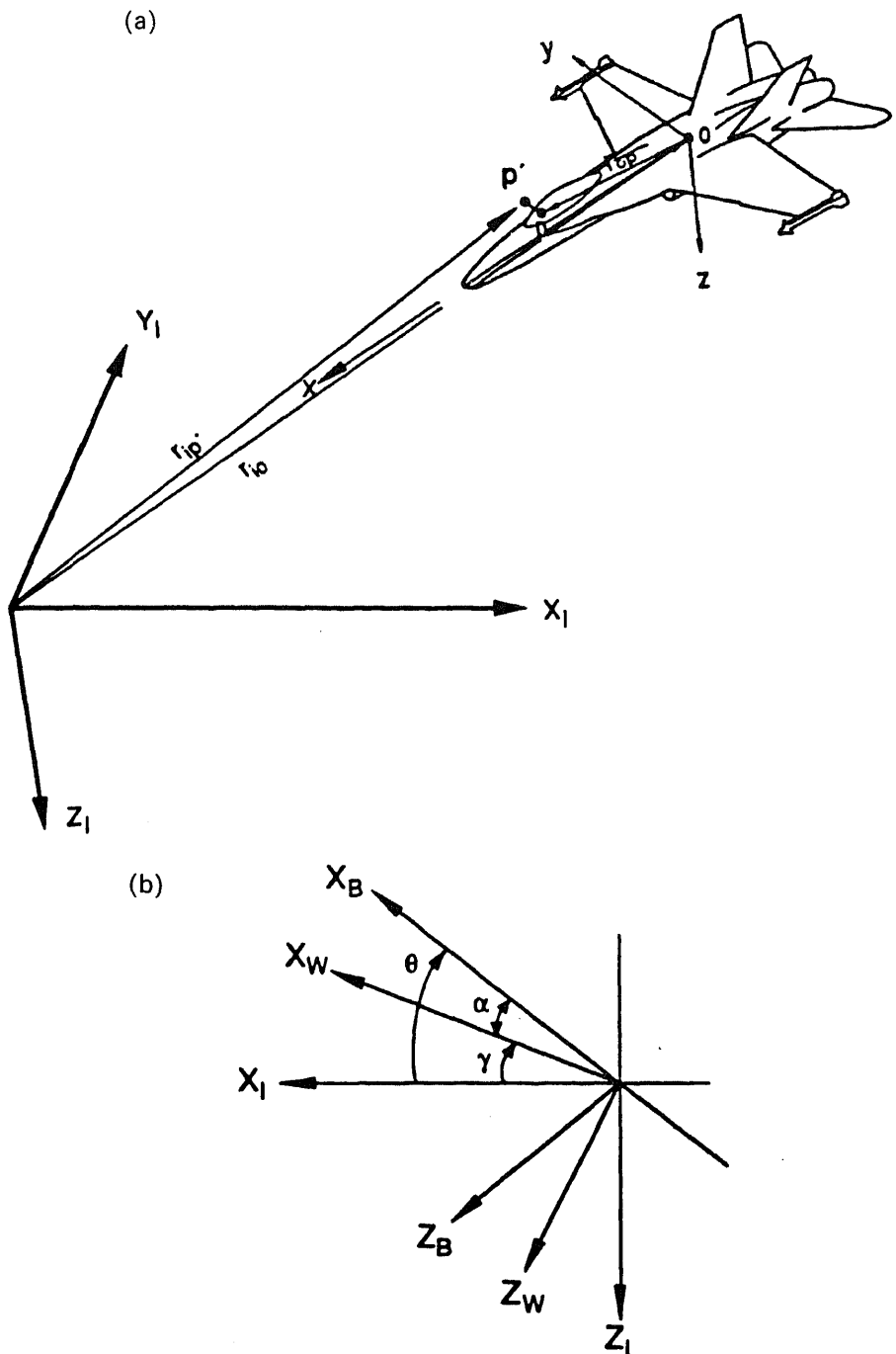
analysis and design cycles so that effective use of the control surfaces may be made to predict the design loads while optimizing aircraft stability, control and performance characteristics. Therefore, the proposed loads analysis method integrates an on-line control law design methodology combining aircraft flexibility and nonlinear aerodynamic forces computed by the CFD methods.

The overall concept of this multidisciplinary analysis approach is depicted in figure 1. The nonlinear aerodynamic forces will be computed by the CFD module. The transient response of the aircraft subject to control stick input will be determined in the aircraft dynamics module. The input and response characteristics data will be analysed in the system identification module to derive the state space matrices. The state space analysis is then employed; (1) to design adaptive flight control laws, (2) to perform flight manoeuvres, and (3) to compute net design loads. The test cases considered in this study demonstrate the computational feasibility of flight manoeuvre analysis using the CFD methods.

Subsequent sections include brief descriptions of: (1) state space formulation of equations of motion of flexible aircraft, (2) implicit acceleration of the CFD method used in the present study, (3) the concept of system identification, and (4) adaptive control law design. A few test cases are included to demonstrate the overall concept of manoeuvre loads analysis.

## 2. Equations of motion of flexible aircraft

The equations of motion of an aircraft may be expressed in any convenient coordinate system. A common practice in flight mechanics is to use a body-fixed coordinate system oriented either along the body axis or the wind axis. Figure 2 shows various axes systems in which the subscripts  $I$ ,  $W$  and  $B$  denote inertial, wind and body-fixed reference coordinates, respectively. The climb angle,  $\gamma$ , is denoted by the angle between



**Figure 2.** Reference axis systems. (a) Body-fixed coordinate system; (b) definition of aerodynamic angles (step input).

the wind axis and the inertial frame of reference, while the aerodynamic angles ( $\alpha, \beta$ ) are denoted by the incidence angles between the wind and the body axes.

A system of equations of motion of a flexible aircraft in the inertial axis system is given by

$$\mathbf{K}\mathbf{r} + \mathbf{M}\ddot{\mathbf{r}} + \mathbf{F}(\dot{\mathbf{r}}, \ddot{\mathbf{r}}, t) = 0 \quad (1)$$

where  $\mathbf{r}$  is a displacement vector,  $\mathbf{K}$  is a stiffness matrix,  $\mathbf{M}$  is a mass matrix and  $\mathbf{F}$  is a vector of aerodynamic and other external forces. The computation of stiffness and mass matrices has by now become a routine procedure in finite element methods (FEM). This generalized analysis procedure in structural mechanics finds its origin in Argyris' work (Argyris & Dunne 1947, 1949; Argyris 1953; Argyris & Kelsey 1960). Hence, there is no need for further discussions regarding the computation of stiffness and mass matrices. On the other hand, the computational aspects of the nonlinear aerodynamics using CFD methods have not yet reached the degree of maturity which has been achieved by the community of structural engineers. Nevertheless, efforts are being made at an ever increasing stride to achieve that goal. We will touch upon this topic a little later.

Equation (1) represents several hundred or even thousands of degrees of freedom. But, if the structure is linear it may be replaced by a fewer modal degrees of freedom,  $\boldsymbol{\eta} = \{\boldsymbol{\eta}_r, \boldsymbol{\eta}_e\}$ , in which the displacement vector  $\mathbf{r}$  will be represented by a linear combination of rigid body modes,  $\Phi_r$ , and the vibration modes,  $\Phi_e$ . Then, the generalized system of equations may be given by (Appa 1991; Argyris & Mlejnek 1991; Argyris *et al* 1995)

$$\bar{\mathbf{K}}\boldsymbol{\eta} + \bar{\mathbf{C}}\dot{\boldsymbol{\eta}} + \bar{\mathbf{M}}\ddot{\boldsymbol{\eta}} + \bar{\mathbf{F}}(\dot{\boldsymbol{\eta}}, \ddot{\boldsymbol{\eta}}, t) = 0. \quad (2)$$

This is a second-order differential equation in time domain. But one can reduce this to a first-order system by replacing the independent variables by

$$\mathbf{x} = \{\boldsymbol{\eta} \dot{\boldsymbol{\eta}}\}, \quad (3)$$

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} = \mathbf{F}(\mathbf{x}, \mathbf{u}, t). \quad (4)$$

The output or the measured signals of the plant may be expressed as a linear combination of  $\mathbf{x}$  and  $\dot{\mathbf{x}}$ , i.e.

$$\mathbf{z} = \mathbf{H}_1\mathbf{x} + \mathbf{H}_2\dot{\mathbf{x}} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \quad (5)$$

where  $\mathbf{u}$  denotes the control surface motion, and  $\mathbf{C}$  and  $\mathbf{D}$  are the measurement matrices.

The system of equations (4) can be integrated analytically provided the eigenvalues of the state matrix  $\mathbf{A}$  lie in the left half of the  $s(= \sigma + i\omega)$  plane with  $\sigma < 0$ . Otherwise, there is a need to find a feedback gain matrix which helps to shift the roots from the right half to the left half of the  $s$ -plane. We will discuss this issue in subsequent sections.

### 3. Computational aerodynamics: CFD

In the last two decades highly competitive numerical schemes have been developed to predict the nonlinear pressure distributions on complex configurations. Hirsch, in

his text books on computational fluid dynamics, discusses a number of numerical schemes developed to date and gives an extensive list of references (Hirsch 1990). These numerical schemes may broadly be classified as those based on structured or unstructured grid models. The finite difference schemes are well suited for structured grid models, while the finite element and the finite volume schemes are applicable for either model. The latter schemes show some potential for generalization of the computational algorithm such as the one used in structural mechanics. Hence, in the present study we have chosen a finite element-based CFD scheme (Argyris *et al* 1989).

The Euler equations of fluid motion, using the notations of Argyris *et al* (1989), may be written as

$$\frac{\partial \mathbf{V}}{\partial t} + \sum_i \frac{\partial \mathbf{f}_i}{\partial x_i} = 0, \quad i = 1, 2, 3, \quad (6)$$

where

$$\mathbf{V} = \{\rho, \rho u_1, \rho u_2, \rho u_3, \rho \varepsilon\} \quad (7)$$

is a vector of independent variables and the flux in the  $i$ th coordinate direction is

$$\mathbf{f}_i = u_i \mathbf{V} + \mathbf{e}_{i+1} p + \mathbf{e}_5 u_i p, \quad i = 1, 2, 3, \quad (8)$$

in which, for example,  $\mathbf{e}_2$  denotes

$$\mathbf{e}_2 = \{0 \ 1 \ 0 \ 0 \ 0\}. \quad (9)$$

The gradient of the flux may be written as

$$\frac{\partial \mathbf{f}_i}{\partial x_i} = \frac{\partial \mathbf{f}_i}{\partial \mathbf{V}} \frac{\partial \mathbf{V}}{\partial x_i} = \mathbf{A}_i \frac{\partial \mathbf{V}}{\partial x_i} \quad (10)$$

where  $\mathbf{A}_i$  is the Jacobian of the flux,  $\mathbf{f}_i$ , in the coordinate direction  $x_i$ . The eigenvalues of this matrix are given by

$$\lambda_k^i = u_i, \quad k = 1, 2, 3 \text{ (for the entropy wave)}, \quad (11)$$

$$\lambda_4^i = u_i + a, \quad \text{(for the forward pressure wave)}, \quad (12)$$

$$\lambda_5^i = u_i - a, \quad \text{(for the backward pressure wave)}. \quad (13)$$

The signs of these eigenvalues denote the direction in which a specific wave moves. Hence, the gradient of the flux  $\mathbf{f}_i$ , must be evaluated in the upwind direction. Further, the Jacobian matrix  $\mathbf{A}_i$ , may be expressed in terms of its eigenvalues and eigenvectors as

$$\mathbf{A}_i = \mathbf{P}_i (\Lambda_i^+ + \Lambda_i^-) \mathbf{P}_i^{-1} = \mathbf{A}_i^+ + \mathbf{A}_i^- \quad (14)$$

where  $\Lambda_i^+$  and  $\Lambda_i^-$  denote diagonal matrices consisting of positive and negative eigenvalues, respectively, and  $\mathbf{P}_i$  is a square matrix of eigenvectors of  $\mathbf{A}_i$ . Then from (10) assuming that  $\mathbf{A}_i$  is constant within an element, we obtain

$$\frac{\partial \mathbf{f}_i}{\partial x_i} = \mathbf{A}_i^+ \frac{\partial^- \mathbf{V}}{\partial x_i} + \mathbf{A}_i^- \frac{\partial^+ \mathbf{V}}{\partial x_i}, \quad (15)$$



where  $(\partial^+/\partial x_i)$  ( ) and  $(\partial^-/\partial x_i)$  ( ) denote upwind differencing depending on the sign of the eigenvalue  $\lambda_i$ .

But we know that the finite element schemes (also finite volume schemes) permit only centre differencing. Hence, it is necessary to express the upwind difference operators in terms of the centre difference operators. This can be achieved by the use of the Taylor series expansion. Thus, using the first two terms in series we obtain

$$\frac{\partial^+ \mathbf{V}}{\partial x_i} = \frac{\partial \mathbf{V}}{\partial x_i} + \sigma \delta_i \left( \frac{\partial \mathbf{V}}{\partial x_i} \right) + \frac{1}{2} \sigma^2 \delta_i^2 \left( \frac{\partial \mathbf{V}}{\partial x_i} \right), \quad 0 < \sigma < 1, \quad (16a)$$

$$\frac{\partial^- \mathbf{V}}{\partial x_i} = \frac{\partial \mathbf{V}}{\partial x_i} - \sigma \delta_i \left( \frac{\partial \mathbf{V}}{\partial x_i} \right) + \frac{1}{2} \sigma^2 \delta_i^2 \left( \frac{\partial \mathbf{V}}{\partial x_i} \right), \quad 0 < \sigma < 1, \quad (16b)$$

where  $\delta_i$  ( ) and  $\delta_i^2$  ( ) denote undivided first- and second-order difference operators. Substituting (16) into (15), and simplifying the algebra, we get

$$\partial \mathbf{f}_i / \partial x_i = \partial \mathbf{f}_i^{**} / \partial x_i, \quad (17)$$

where the upwind flux,  $\mathbf{f}_i^{**}$ , is given by

$$\mathbf{f}_i^{**} = \mathbf{f}_i^*(\mathbf{V}_i^*) - \sigma |\mathbf{A}_i| \delta_i \mathbf{V}, \quad (18)$$

$$\mathbf{V}_i^* = \mathbf{V} + \frac{1}{2} \sigma^2 \delta_i^2 \mathbf{V}, \quad (19)$$

$$|\mathbf{A}_i| = \mathbf{P}_i (\Lambda_i^+ + |\Lambda_i^-|) \mathbf{P}_i^{-1}. \quad (20)$$

Thus, we require only central differencing which can be accomplished in the finite element scheme. The original ICA-CFD code has been modified to compute the upwind flux gradients. This eliminates the need for the artificial viscosity terms, which are arbitrary and undesirable in unsteady aerodynamic analysis consisting of low and high frequency spectra. Additional discussions relevant to the numerical computation of the upwind flux may be found in Argyris *et al* (1995).

### 3.1 Moving boundary conditions

The solid boundary may undergo a set of rigid body motions,  $\eta_r$ , and elastic motions,  $\eta_e$ , as discussed earlier. The rigid body motions contribute to aerodynamic attitudes resulting from lateral motions  $(\dot{x}, \dot{y}, \dot{z})$  (e.g. angle of attack,  $\alpha$ , and sideslip,  $\beta$ ) and angular rates  $(\dot{\phi}, \dot{\theta}, \dot{\psi})$  (roll rate,  $p$ , pitch rate,  $q$ , and yaw rate,  $r$ ). These are the relative velocities with respect to the infinite volume of fluid. Hence, the flow variables at all points must be updated. Thus, at each fluid node,  $n$ , we get

$$\mathbf{v}' = \mathbf{T}(\Delta\alpha, \Delta\beta) \mathbf{v} + [\Phi_r]_n \boldsymbol{\theta}, \quad (21)$$

where

$$\mathbf{v} = \{u_1, u_2, u_3\} \text{ are nodal velocities,} \quad (22)$$

$$\boldsymbol{\theta} = \{dp, dq, dr\} \text{ are incremental angular velocities,} \quad (23)$$

$\mathbf{T}(\Delta\alpha, \Delta\beta)$  is a transformation matrix based on the incremental incidences,

and

$\Phi_r$  is a rigid body modal matrix at node  $n$ .

On the solid surface, the incremental velocity vector is given by the incremental change of slopes  $(\Delta\phi, \Delta\theta, \Delta\psi)$  arising from the elastic deformations as well as the vibration,  $\dot{\eta}_e$ , i.e.

$$\Delta v = T[\Delta\phi, \Delta\theta, \Delta\psi]v + [\Phi_e]\dot{\eta}_e. \quad (24)$$

These modifications have been implemented in the ICA-CFD code.

### 3.2 Implicit acceleration scheme

The original ICA-CFD solution algorithm has been replaced by a quasi-implicit scheme discussed in Argyris *et al* (1995). Equation (6) in the context of a finite element scheme may be rewritten as

$$M(\partial V/\partial t) + R = 0, \quad (25)$$

where  $M$  is the mass-like matrix (usually a lumped diagonal matrix) and  $R$  is the residual flux vector. After multiplying by  $M^{-1}$  we get

$$(\partial V/\partial t) + \bar{R} = 0. \quad (26)$$

The normalized flux,  $\bar{R}$  can be expressed as

$$\bar{R} = \bar{R}_0 + \left[ \frac{\partial \bar{R}}{\partial V} \right] \Delta V, \quad (27)$$

$(\partial \bar{R}/\partial V)$  is a Jacobian matrix, which can be approximated as a banded matrix consisting of 3 to 9 off-diagonal elements. The elements of  $(\partial \bar{R}/\partial V)$  can be computed using the concept of system identification. The computational details may be found in Argyris *et al* (1995). The solution algorithm is then given by

$$\Delta V = \Delta t [I + \Delta t (\partial \bar{R}/\partial V)]^{-1} \bar{R}_0. \quad (28)$$

The stability of the system of equations increases with increasing bandwidth. In the present analysis three diagonal elements were used with the global Courant–Friedrichs–Lewy (CFL) number  $\leq 4.0$ . A two-step solution is recommended for nonstationary solutions.

## 4. System identification

System identification is a recursive on-line estimation procedure employed to construct equivalent mathematical models of dynamic systems using input and output sequences. This methodology has been widely applied in advanced communications, space flights and industrial automation. Here we intend to apply this scheme to determine the linearized dynamic models,  $A$  and  $B$ , as proposed in §2 (i.e. (4)).

The usual approach is to represent the differential equations in terms of  $N$  previously

known records, such as

$$\mathbf{y}_{n-1} = \sum_{n=1}^N (\mathbf{A}_n \mathbf{y}_n + \mathbf{B}_n \mathbf{u}_n), \quad (29)$$

and solve for the  $\mathbf{A}_n$  and  $\mathbf{B}_n$  matrices (Balakrishnan 1968; Hsia 1977; Åström & Wittenmark 1989). The resulting dynamical model and the feedback gain matrix will be  $N$  times larger than the original system (Pak & Friedman 1991). This means extensive computer resources are required to design an on-line adaptive control law. To minimize the cost of computation Argyris *et al* (1995) employ an alternate method which requires only one sampling from an earlier record. In other words, the number of equations in the estimation process is the same as in the original model. This is a significant reduction in the overhead cost of designing a feedback control system. A brief account of this procedure is as follows:

Let

$$\dot{\mathbf{x}} = \mathbf{F}^*(\mathbf{x}, \mathbf{u}, t), \quad (30)$$

where

$\mathbf{F}^*$  is a vector of normalized aerodynamic forces;

$\mathbf{x}$  and  $\mathbf{u}$  represent the aerodynamic angles ( $\alpha, \beta$ ), angular velocities ( $p, q, r$ ) and the control surface motion (these are the inputs to the CFD module);

and

$\dot{\mathbf{x}}$  represents the acceleration of the system as an output.

The desired mathematical model may then be linearized to read as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{f}_0. \quad (31)$$

This equation, in the notations of autoregression scheme, may be rewritten as

$$\mathbf{y} = \boldsymbol{\phi}^T \boldsymbol{\theta}, \quad (32)$$

where

$$\mathbf{y} = \dot{\mathbf{x}}^T, \quad (33)$$

$$\boldsymbol{\phi} = \{\mathbf{x}, \mathbf{u}\}, \quad (34)$$

$$\boldsymbol{\theta} = \begin{bmatrix} \mathbf{A}^T \\ \mathbf{B}^T \end{bmatrix}. \quad (35)$$

Since  $\mathbf{y}$  and  $\boldsymbol{\phi}$  are known at various time intervals, the autoregression procedure can be used to solve for  $\boldsymbol{\theta}$ .

## 5. Adaptive feedback control design

Modern aircraft designed with relaxed static stability require digital flight control systems to provide stability and control at all flight conditions. Moreover, if the pressure distributions as computed by the CFD methods are nonlinear functions of the flight parameters then the dynamic model will be changing with time. For these reasons it is necessary that an adaptive control law based on the current estimate of aerodynamic derivatives must be implemented. In a previous section we described

a procedure for estimating the dynamic characteristics of a system using the concept of system identification. We will now describe a design procedure which utilizes this information to design a feedback control system.

Let us consider a system of state space equations given by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad (36)$$

and the corresponding output equations by

$$\mathbf{z} = \mathbf{H}_1\mathbf{x} + \mathbf{H}_2\dot{\mathbf{x}} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \quad (37)$$

where  $\mathbf{H}_1$  and  $\mathbf{H}_2$  are the output transformation matrices. The  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$  and  $\mathbf{D}$  matrices constitute what is generally known as the quadruplets of a set of state space equations. If  $\mathbf{z}_T$  is a vector of target output of a system, the performance index function may be written as

$$J = h_T + \int_t^{t_f} \left[ \frac{1}{2}\boldsymbol{\varepsilon}^T \mathbf{Q}\boldsymbol{\varepsilon} + \frac{1}{2}(\mathbf{u} - \mathbf{u}_c)^T \mathbf{R}(\mathbf{u} - \mathbf{u}_c) + \mathbf{p}^T(\mathbf{F} - \dot{\mathbf{x}}) \right] dt, \quad (38)$$

where

$\mathbf{u}_c$  is the pilot command,

$\mathbf{p}$  is a vector of Lagrangian coefficients,

$\boldsymbol{\varepsilon} = (\mathbf{z} - \mathbf{z}_T)$  is an error vector,

$\mathbf{Q}$  is a symmetric weighting matrix,

$\mathbf{R}$  is a symmetric and positive definite matrix, (39)

and

$$h_T = \frac{1}{2}\boldsymbol{\varepsilon}_T^T \mathbf{Q}\boldsymbol{\varepsilon}_T \text{ is the terminal performance index at } t_f. \quad (40)$$

The designer selects the weighting matrices,  $\mathbf{Q}$  and  $\mathbf{R}$ , to satisfy the stability requirements. Parameters,  $\mathbf{x}$ ,  $\mathbf{u}$  and  $\mathbf{p}$  must be determined to minimize the performance index,  $J$ . Hence, the variation of  $J$  with respect to  $\mathbf{x}$ ,  $\dot{\mathbf{x}}$ ,  $\mathbf{u}$  and  $\mathbf{p}$ , yields a system of equations for  $\mathbf{x}(t)$ ,  $\mathbf{p}(t)$  and  $\mathbf{u}(t)$ ,

$$\dot{\boldsymbol{\xi}}(t) = \mathbf{H}\boldsymbol{\xi} + \mathcal{T} \quad (41)$$

and

$$\mathbf{u} = -\mathcal{R}^{-1}[\mathbf{B}^T \mathbf{p}(t) + \mathbf{D}^T \mathbf{Q}(\mathbf{C}\mathbf{x}(t) - \mathbf{z}_T)] \quad (42)$$

where

$$\boldsymbol{\xi} = \{\mathbf{x}, \mathbf{p}\}, \quad (43)$$

$$\mathbf{H} = \begin{bmatrix} \mathcal{A} & -\mathcal{B} \\ -\tilde{\mathcal{Q}} & -\mathcal{A}^T \end{bmatrix} \text{ (Hamiltonian matrix),} \quad (44)$$

$$\mathcal{T} = \begin{Bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{Bmatrix} \quad (45)$$

in which

$$\mathcal{R} = \mathbf{R} + \mathbf{D}^T \mathbf{Q} \mathbf{D}, \quad (46)$$

$$\mathcal{A} = \mathbf{A} - \mathbf{B}\mathcal{R}^{-1}\mathbf{D}^T \mathbf{Q} \mathbf{C}, \quad (47)$$

$$\mathcal{B} = \mathbf{B}\mathcal{R}^{-1}\mathbf{B}^T, \quad (48)$$

$$\tilde{\mathbf{Q}} = \mathbf{C}^T[\mathbf{Q} - \mathbf{Q}\mathbf{D}\mathcal{R}^{-1}\mathbf{D}^T \mathbf{Q}]\mathbf{C}, \quad (49)$$

$$\mathbf{F}_1 = \mathbf{B}\mathcal{R}^{-1}\mathbf{D}^T \mathbf{Q} \mathbf{z}_T, \quad (50)$$

$$\mathbf{F}_2 = \mathbf{C}^T[\mathbf{Q} - \mathbf{Q}\mathbf{D}\mathcal{R}^{-1}\mathbf{D}^T \mathbf{Q}]\mathbf{z}_T. \quad (51)$$

The resulting end conditions are given by

$$\mathbf{x}(t_0) = \mathbf{x}_0, \quad (52)$$

$$\mathbf{p}(t_f) = [\mathbf{C}\mathbf{x}(t_f) + \mathbf{D}\mathbf{u}(t_f) - \mathbf{z}_T]. \quad (53)$$

This is a two-point boundary value problem. Its solution is more complicated than the usual one-point boundary value problem. A systematic solution procedure is given by Kirk (1970). For the sake of completeness we briefly summarize his procedure. The solution to (41) may be written as

$$\xi(t_f) = [\exp(\mathbf{H}(t_f - t))] \xi(t) + \mathbf{H}^{-1}[\exp(\mathbf{H}(t_f - t)) - \mathbf{I}] \mathcal{T}. \quad (54)$$

Using the terminal end condition given by (53), one can solve for the Lagrangian coefficients,  $\mathbf{p}$ , in the following form:

$$\mathbf{p}(t) = \mathbf{K}(t)\mathbf{x}(t) + \mathbf{z}'(t). \quad (55)$$

Substituting (55) into (42) the adaptive feedback signal reduces to

$$\mathbf{u} = \bar{\mathcal{X}}(t)\mathbf{x}(t) + \bar{\mathbf{z}}(t) \quad (56)$$

and the adaptive control law is given by  $\bar{\mathcal{X}}(t)$ .

If the target values,  $\mathbf{z}_T$  were set to zero, then the above tracking problem reduces to that of a regulator. In that case  $\bar{\mathbf{z}}(t) = 0$ .

Finally, substituting for  $\mathbf{u}$  in (36) we obtain a set of stable plant equations

$$\dot{\mathbf{x}} = [\mathbf{A} + \mathbf{B}\bar{\mathcal{X}}(t)]\mathbf{x} = \bar{\mathbf{A}}\mathbf{x}. \quad (57)$$

The designer may choose  $\mathbf{Q}$  and  $\mathbf{R}$  such that the real components of the roots of  $\bar{\mathbf{A}}$  are negative (i.e., the roots lie in the left half of the  $s$ -plane).

## 6. Aircraft manoeuvre analysis

In this section we specialize the equations of motion discussed in the previous sections to compute the design loads of a manoeuvring aircraft. The inertial accelerations arising from vibration are negligibly small compared to the manoeuvring accelerations and hence the acceleration terms in (2) which involve the elastic deformation terms,  $\ddot{\eta}_e$  and  $\dot{\eta}_e$ , omitted. Then we obtain a system of equations of motion in rigid body modes including the flexibility effects of the aircraft structure. After some simplifications the final manoeuvre equation (2) reduces to the following form:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, \mathbf{u}, t), \quad (58)$$

in which the state variables are given by

$$\mathbf{x} = \{V, \gamma, q, \theta\}, \quad (59)$$

and the control variables are

$$\mathbf{u} = \{\delta_1, \delta_2, \delta T_1, \delta T_2\}, \quad (60)$$

in which  $V$  is the aircraft velocity along the flight path,  $\delta_i$  is the control surface deflection and  $\delta T_i$  is the thrust control. The column vector  $\mathbf{F}$  represents the mass normalized forces along and normal to the wind axis and the pitching moment about its mass centre. The last element in  $\mathbf{F}$  is  $F_4 = q = \dot{\theta}$  and represents an integrator to obtain  $\theta$ . The aerodynamic forces will be computed iteratively using the deformed shape of the aircraft given by

$$\eta_e = -[\mathbf{K}_e]^{-1} \mathbf{F}_e. \quad (61)$$

The generalized force vector  $\mathbf{F}_e$  represents the net forces (aerodynamic-inertia (centrifugal)) on the aircraft.

The system of equations in (58) may be solved iteratively in the following steps:

- (1) compute  $\dot{\mathbf{x}}$  using initial values of  $\mathbf{x}$  and  $\mathbf{u}$ ,
- (2) integrate  $\dot{\mathbf{x}}$  to get  $\mathbf{x}$ ,
- (3) perform system identification to determine the state space matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,
- (4) compute the feedback gain matrix and the feedback signal  $\mathbf{u}$ ,
- (5) compute the target error,  $\varepsilon = \mathbf{z} - \mathbf{z}_T$ ,
- (6) if  $|\varepsilon|$  less than or equal to a specified value, end the computation,
- (7) otherwise, repeat steps 1 through 6.

Finally, integrate the net forces to yield section design loads such as shear, bending and torsion.

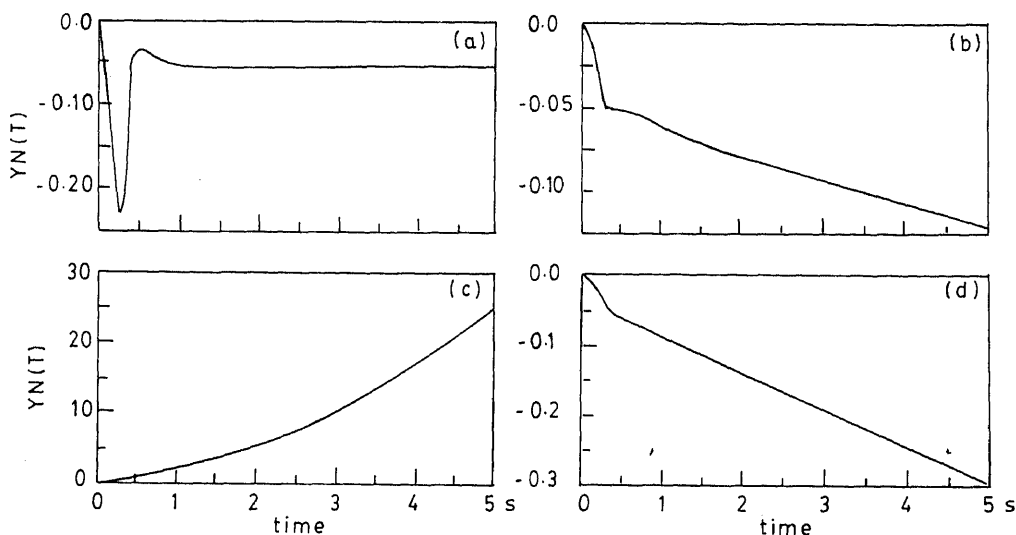
## 7. Discussion of results

In the previous sections we discussed an analytical method unifying the CFD algorithm to compute flight loads and aeroelastic stability boundaries. Some preliminary test cases have been performed to validate the concept in the following areas:

- (a) system identification;
- (b) on-line adaptive control design;
- (c) aircraft manoeuvre analysis;
- (d) static aeroelastic effects and aeroservoelastic stability analysis.

### 7.1 System identification and control design

Two test cases have been selected to verify the accuracy and the reliability of these two modules. The first example represents the landing approach of the F-16 model reported by Rynaski (1982). The state variables are  $\mathbf{x} = \{q, V, \alpha, \theta\}$  and the control variable is the elevator deflection  $\mathbf{u} = \{\delta_e\}$ . The corresponding state space matrices  $\mathbf{A}$  and  $\mathbf{B}$  are given in table 1. For a unit-step input the plant response,  $\dot{\mathbf{x}}$ , was computed and integrated to obtain  $\mathbf{x}$ . Then, the autoregression module was used to estimate the plant matrices,  $\bar{\mathbf{A}}$  and  $\bar{\mathbf{B}}$ . Subsequently, using  $\bar{\mathbf{A}}$  and  $\bar{\mathbf{B}}$  the feedback control gain matrix,  $\mathbf{K}$ , was computed. In addition the eigenvalues of the open-loop and the closed-loop plant matrices were calculated. The estimate matrices after fifty time-steps are seen to be in excellent agreement with the original data. Since the F-16 is a statically relaxed aircraft one of its roots is unstable. But, after closing the loop through the feedback control all the roots move to the left half of the  $s$ -plane. The closed-loop



**Figure 3.** F-16 landing approach manoeuvre model (Rynaski 1982). (a) Pitch rate,  $Q$ ; (b)  $\alpha$ ; (c) speed change,  $DV$ ; (d) pitch angle.

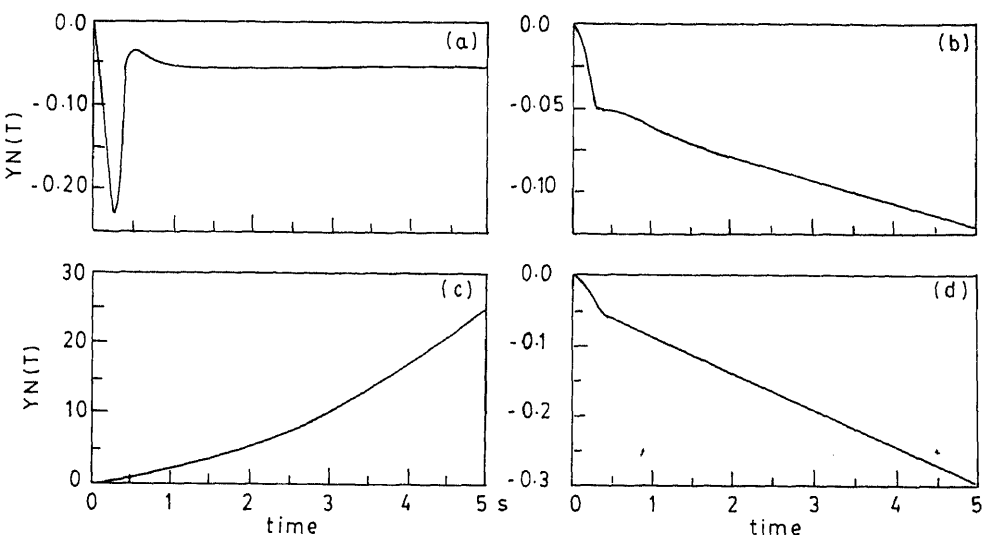
response characteristics of the aircraft due to a unit-step input are shown in figure 3. The system responds very rapidly to the elevator command to gain the desired lift during a landing manoeuvre.

The second example represents the fuselage pitch pointing of the F-16 model. The state variables in this case are  $x = \{\theta, q, \alpha\}$ , and the control variables are  $u = \{\delta_f(\text{flap}), \delta_e(\text{elevator})\}$ . The plant matrices as reported by Porter & Bradshaw (1981) are given in table 2. Once again the estimated plant matrices are in excellent agreement with the original data. As in the previous example the unstable roots move from the right half of the  $s$ -plane to the stable zone in the left half plane. The closed-loop response characteristics of the aircraft are shown in figure 4. Any desired response characteristics can be achieved by selecting appropriate weighting coefficients,  $Q$  and  $R$ .

These two examples and many others which are not reported here justify the use of the single sampling approach suggested in this study to estimate the plant matrices with reasonable accuracy. In addition, the on-line control design procedure used here is able to stabilize the system whose dynamic characteristics are practically unknown to begin with.

## 7.2 Symmetric flight manoeuvre analysis

To demonstrate the simulation of the manoeuvre analysis using the CFD scheme we have selected a simple delta wing with a cranked leading edge as shown in figure 5. Two control surfaces, one inboard and one outboard, were used to trim and control the vehicle. The gross weight and the pitching moment of inertia of the vehicle were  $1.55 \times 10^6 \text{ N}$  and  $3.019 \times 10^6 \text{ kg} \cdot \text{m}^2$ . The centre of gravity was located at 60% of the mid chord. For 1- $g$  level flight at Mach 0.7, the analysis started with an initial angle to attack of  $3^\circ$ . First, 50 time-step solutions were performed just to establish the flow field around the aircraft. Then the balancing of the aircraft started with the feedback control system turned on. The state space matrices were continuously estimated



**Figure 3.** F-16 landing approach manoeuvre model (Rynaski 1982). (a) Pitch rate,  $Q$ ; (b)  $\alpha$ ; (c) speed change,  $DV$ ; (d) pitch angle.

response characteristics of the aircraft due to a unit-step input are shown in figure 3. The system responds very rapidly to the elevator command to gain the desired lift during a landing manoeuvre.

The second example represents the fuselage pitch pointing of the F-16 model. The state variables in this case are  $\mathbf{x} = \{\theta, q, \alpha\}$ , and the control variables are  $\mathbf{u} = \{\delta_f(\text{flap}), \delta_e(\text{elevator})\}$ . The plant matrices as reported by Porter & Bradshaw (1981) are given in table 2. Once again the estimated plant matrices are in excellent agreement with the original data. As in the previous example the unstable roots move from the right half of the  $s$ -plane to the stable zone in the left half plane. The closed-loop response characteristics of the aircraft are shown in figure 4. Any desired response characteristics can be achieved by selecting appropriate weighting coefficients,  $\mathbf{Q}$  and  $\mathbf{R}$ .

These two examples and many others which are not reported here justify the use of the single sampling approach suggested in this study to estimate the plant matrices with reasonable accuracy. In addition, the on-line control design procedure used here is able to stabilize the system whose dynamic characteristics are practically unknown to begin with.

## 7.2 Symmetric flight manoeuvre analysis

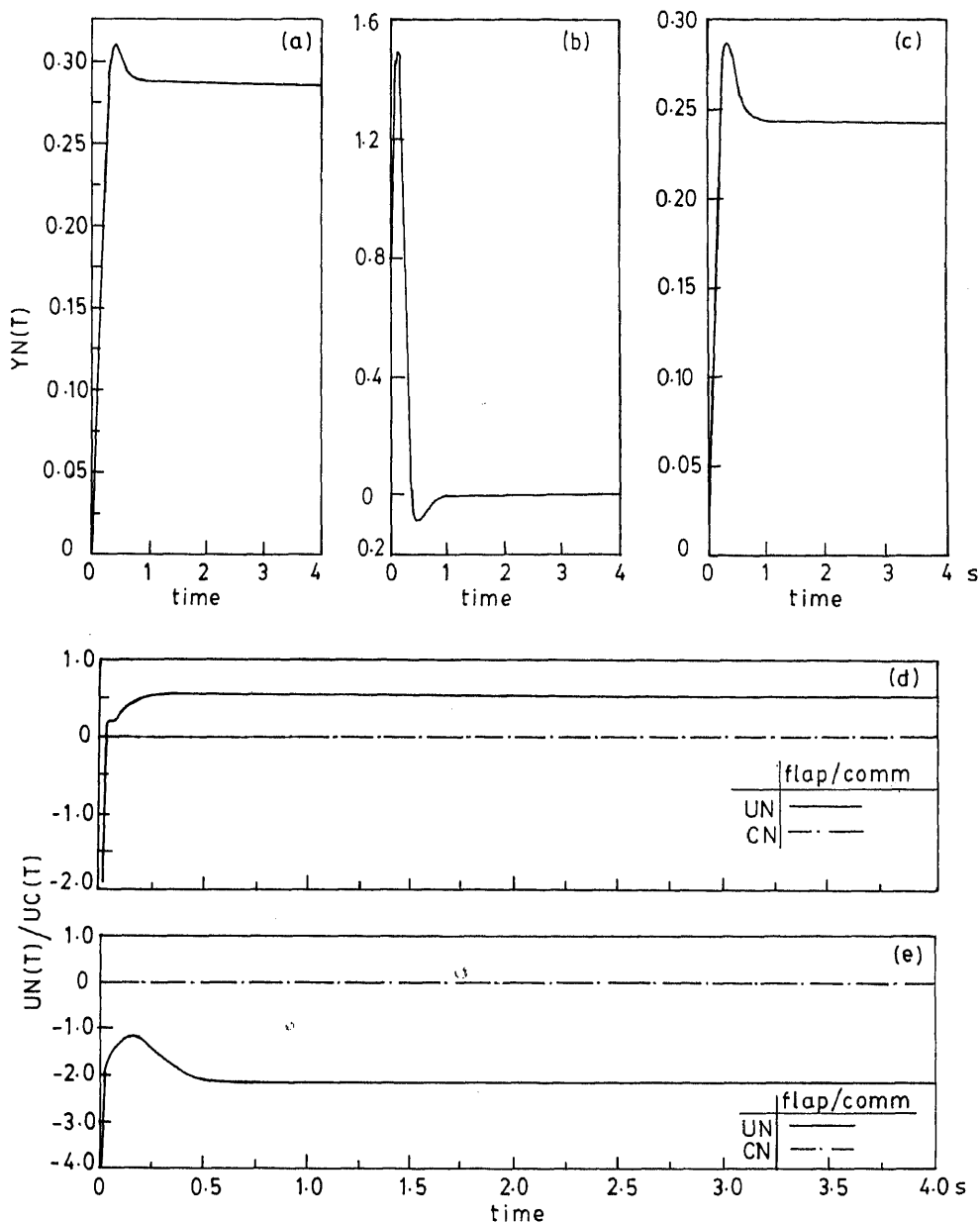
To demonstrate the simulation of the manoeuvre analysis using the CFD scheme we have selected a simple delta wing with a cranked leading edge as shown in figure 5. Two control surfaces, one inboard and one outboard, were used to trim and control the vehicle. The gross weight and the pitching moment of inertia of the vehicle were  $1.55 \times 10^6$  N and  $3.019 \times 10^6$  kg·m<sup>2</sup>. The centre of gravity was located at 60% of the mid chord. For 1- $g$  level flight at Mach 0.7, the analysis started with an initial angle to attack of 3°. First, 50 time-step solutions were performed just to establish the flow field around the aircraft. Then the balancing of the aircraft started with the feedback control system turned on. The state space matrices were continuously estimated



Table 2. F-16 fuselage pitch pointing manoeuvre.

Parameters	$\theta$	$q$	$\alpha$	$\delta_f$	$\delta_e$
Open-loop roots	<i>Original A-matrix</i>				
	-0.0000E+00	0.1000E+01	0.0000E+00	0.0000E+00	0.0000E+00
	0.0000E+00	-0.1779E+01	0.6953E+01	-0.2897E+02	-0.6453E+01
	0.0000E+00	0.9890E+00	-0.1660E+01	-0.1740E+00	-0.2300E+00
	<i>Real</i>				
	0.3683E-04		0.0000E+00		
	-0.4361E+01		0.0000E+00		
	0.9027E+00		0.0000E+00		
	<i>Imaginary</i>				
Closed-loop $\bar{A}$ -matrix $A = A - BK$	<i>Estimated A-matrix</i>				
	-0.6697E-05	0.1002E+01	-0.6899E-02	0.2887E-01	0.6433E-02
	0.7663E-03	-0.1789E+01	0.6976E+01	-0.2902E+02	-0.6463E+01
	-0.2075E-03	0.9924E+00	-0.1669E+01	-0.1458E+00	-0.2240E+00
	0.2027E+00	0.1018E+01	-0.8939E-01		
	-0.2038E+03	-0.1789E+02	0.8993E+02		
	-0.5006E+00	0.8025E+00	-0.2462E+01		
	<i>Real</i>				
	-0.7010E+01		-0.6329E+01		
	-0.7010E+01		0.6329E+01		
	-0.6131E+01		-0.0000E+00		
Closed-loop roots	<i>Estimated B-matrix</i>				
	<i>Imaginary</i>				

Reference: Porter &amp; Bradshaw (1981)



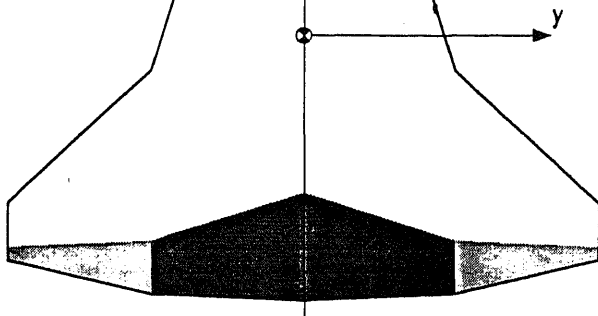
**Figure 4.** Fuselage pitch pointing manoeuvre model (Porter & Bradshaw 1981): (a)  $\theta$ ; (b) pitch rate,  $Q$ ; (c)  $\alpha$ ; (d) flaperon (degrees); (e) elevator (degrees).

and the feedback gain matrices were computed. The required load factor,  $n_z$ , was used as the terminal target quantity.

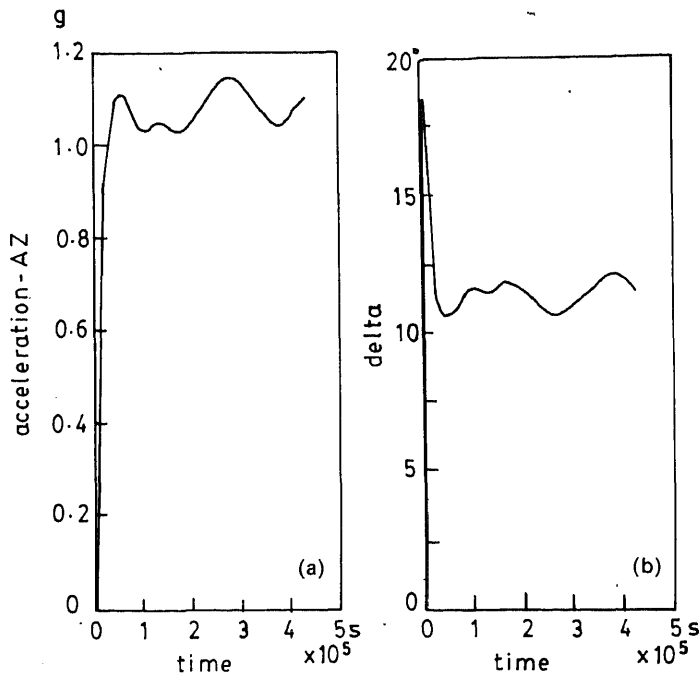
The estimated plant matrices for the open-loop and the closed-loop cases are shown in table 3. The elements of these matrices denote the stability derivatives of the aircraft with respect to  $V$ ,  $\gamma$ ,  $q$ ,  $\theta$  and  $\delta$  parameters. These quantities may also be expressed

Table 3. State space matrices in symmetric manoeuvre for cranked delta wing at  $M = 0.7$ .

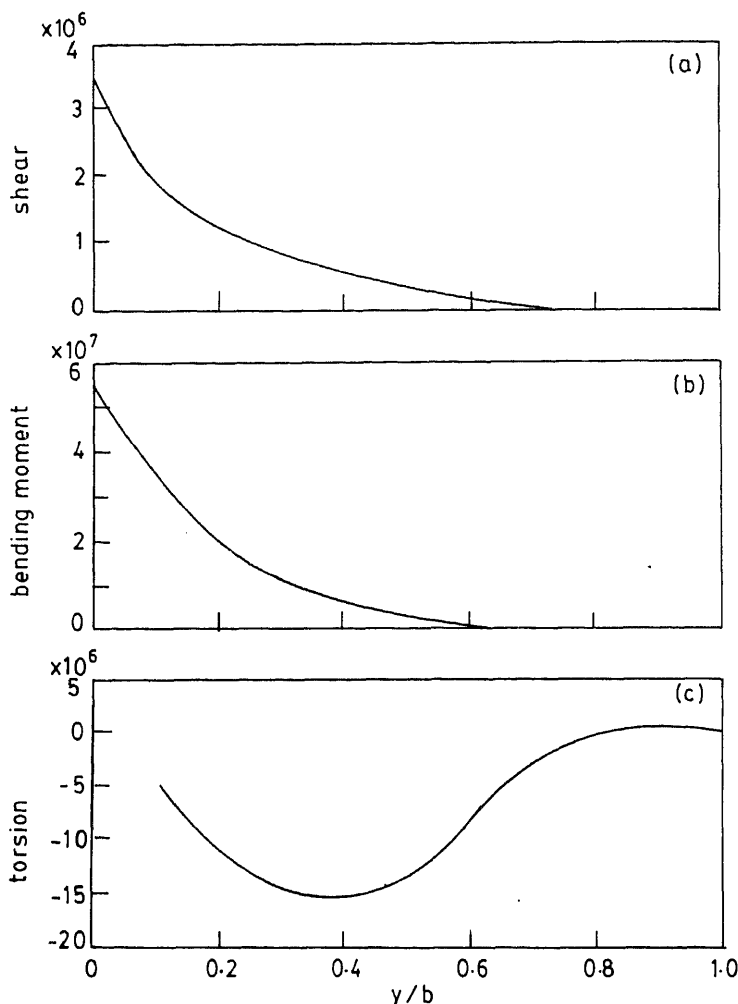
Parameters	$\Delta V$	$\gamma$	$q$	$\theta$	$\delta_{\text{Inboard}}$	$\delta_{\text{Outboard}}$
Open-loop			<b>A-matrix</b>			<b>B-matrix</b>
	0.1166E-03	-0.1495E+03	0.4313E+03	0.7967E+01	-0.1012E-01	-0.7004E-02
	-0.5112E-04	-0.3910E+01	0.1859E+02	0.1767E+00	-0.2453E-03	-0.2257E-03
	0.9231E-03	0.6123E+02	-0.7447E+03	-0.2831E+01	0.9691E-02	0.1010E-01
	0.6558E-07	-0.6558E-01	-0.9912E+00	0.3965E-02	0.2785E-06	0.1152E-05
Roots		<i>Real</i>	<i>Imaginary</i>			
		-0.7462E+03	0.0000E+00			
		-0.2377E+01	0.0000E+00			
		0.2562E-02	0.0000E+00			
		-0.1495E-02	0.0000E+00			
Closed-loop						
<b>A = A - BK</b>						
	-0.3357E-04	-0.1947E+03	0.4683E+03	0.1039E+02		
	-0.5511E-04	-0.4702E+01	0.1926E+02	0.2177E+00		
	0.1081E-02	0.6475E+02	-0.7509E+03	-0.2870E+01		
	0.7675E-07	-0.1078E-01	-0.9915E+00	0.4021E-02		
Roots		<i>Real</i>	<i>Imaginary</i>			
		-0.7525E+03	0.0000E+00			
		-0.3034E+01	0.0000E+00			
		0.2044E-02	0.0000E+00			
		-0.1450E-02	0.0000E+00			



**Figure 5.** Cranked delta wing with inboard and outboard flaps.



**Figure 6.** 1- $g$  symmetric level flight. (a) Normal acceleration in  $g$  vs time; (b) control is time.



**Figure 7.** Integrated load distributions for 1-g flight at  $M=0.7$ . (a) Shear; (b) bending moment; (c) torsion.

in terms of usual aerodynamic stability derivatives such as:  $C_{L\alpha}$ ,  $C_{M\alpha}$ ,  $C_{Lq}$ ,  $C_{L\delta}$  etc. for rigid as well as flexible aircraft. Figure 6 shows the time history of acceleration and the control surface rotation for 1-g level flight. The corresponding integrated load distributions: shear, bending moment and torsion about a reference axis at the centre of gravity are shown in figure 7. The manoeuvre data presented in this example remain to be verified with other sources for accuracy and reliability.

## 8. Conclusions

This paper summarizes a new multidisciplinary approach to computational aircraft dynamics and loads analysis. System identification has been extensively used to solve the CFD problem and to estimate the state-space matrices, which determine the stability characteristics of the system and to compute adaptive gain matrices. The

principle of multidisciplinary analysis alluded to in this study has other promising areas for application and development. Specifically these include: (1) computation of aerodynamic stability derivatives of rigid and flexible aircraft, (2) aeroservoelastic analysis, and (3) multidisciplinary design and optimization. While the results of the specific examples presented in this study are encouraging, additional validation of the method using practical examples has yet to be addressed.

The computational effort of this project was supported by IBM of the United States of America. The authors sincerely thank Mr Raj Mantha of IBM for his active participation and interest in the progress of the project. Mr Don Rossman of IBM made necessary conversions to port the ICA-CFD code onto the IBM 3090 system. We thank him for this effort. The authors also wish to thank Messrs Juri Kalviste and Don Kesler for their valuable discussions during the development of the adaptive control laws. Finally, the authors' appreciation and thanks go to Ms R Corvese for her skillful and dedicated effort in the integration of the loads module with the CFD code.

## References

- Appa K 1991 Recent advances in manoeuvre loads analysis. *Comput. Meth. Appl. Mech. Eng.* 90: 693-717
- Argyris J 1953 Thermal stress analysis and energy theorems. *A. R. C* 16489
- Argyris J, Appa K, Bühlmeier J 1995 *Texts on computational mechanics: Vol. VI. A survey of aeroservoelasticity* (Amsterdam: North-Holland) (to be published)
- Argyris J, Dunne P C 1947 The general theory of cylindrical and conical tubes under torsion and bending loads. *J. R. Aeronaut. Soc.* 51: 199-269, 757-784, 884-930
- Argyris J, Dunne P C 1949 The general theory of cylindrical and conical tubes under torsion and bending loads. *J. R. Aeronaut. Soc.* 53: 461-483, 558-620
- Argyris J, Kelsey S 1960 *Part II: Energy theorems and structural analysis* (London: Butterworths) (originally published in series of articles in *Aircraft Engineering*, Oct, Nov 1954 and Feb, March, April, May 1985)
- Argyris J, Mlejnek H P 1991 *Texts on computational mechanics. Vol. V. Dynamics of structures* (Amsterdam: North-Holland)
- Argyris J, St. Doltsinis I, Friz H 1989 Hermes space shuttle: Exploration on reentry aerodynamics. *Comput. Meth. Appl. Mech. Eng.* 73: 1-51
- Åström K J, Wittenmark B 1989 *Adaptive control* (Addison-Wesley)
- Balakrishnan A V 1968 A new computing technique in system identification, *J. Computing System Sci.* 2: 102-116
- Hirsch C 1990 *Numerical computation of internal and external flows* (Chishester: John Wiley & Sons) Vols. 1 and 2
- Hsia T C 1977 *System identification: Least square methods* (Lexington, MA: Lexington books)
- Kirk D E 1970 *Optimal control theory: An introduction* (Network Series) (ed.) R W Newcomb, (Englewood Cliffs, NJ: Prentice-Hall)
- Pak C G, Friedmann P P 1991 Transonic adaptive flutter suppression using approximate unsteady time domain aerodynamics. Paper No. AIAA-91-0986
- Porter B, Bradshaw A 1981 *Design of direct digital flight mode control systems for high-performance aircraft* (New York: IEEE)
- Rynaski E G 1982 Flight control system design using robust output observers. *Advances in Guidance and Control System: AGARD Conference proceeding*, No. 321



## Recent progress in dynamics and aeroelasticity

A R UPADHYA and KESHAB PANDA

Aeronautical Development Agency, P B No. 1718, Vimanapura PO,  
Bangalore, 560 017, India

**Abstract.** With the emphasis on higher performance, modern aircraft designs aim at lower structural weight, aerodynamically efficient thinner configurations, and reduced or even negative stability margins augmented by automatic flight control systems. These design considerations lead to highly flexible structural designs with associated problems of dynamic and aeroelastic interactions which need to be considered at the preliminary design stage itself. Advent of advanced composites and active control techniques have given the aircraft designer the freedom to use aeroelastic interactions in an advantageous way. This paper reviews the recent research and development efforts in the areas of aeroelastic tailoring, structural optimisation with aeroelastic constraints and aeroservoelasticity, and applications of the same in practical designs. The developments and applications in India in these areas are also highlighted.

**Keywords.** Aeroelasticity; aerodynamic interactions; aeroelastic tailoring; aeroservoelasticity; structural optimisation.

### 1. Introduction

The general developments in the aerospace industry during the last two decades have called for a fresh look at the role of dynamics and aeroelasticity in aerospace vehicle design. Higher performance and extended mission requirements demand designs with low structural weight, aerodynamically efficient configurations and reduced or even negative stability margins augmented by automatic flight control systems. These conflicting design requirements lead to highly flexible structural designs with associated problems of dynamic, aeroelastic and control interactions which need to be predicted and controlled at the preliminary design stage itself without undue weight penalty. Greater emphasis on the transonic flight regime presents its own special problems of flow unsteadiness and resulting structural response and flutter. Another important development that has a strong impact on design is the increasing use of advanced composite materials in primary structures such as wing and empennage. While composites offer attractive weight savings in view of their higher specific strength and stiffness properties, they also present the problem of anisotropy and resultant coupled deformation behaviours. Research studies in the past two decades have shown how the directional properties of composites could be put to advantageous use



through aeroelastic tailoring techniques to obtain improved aeroelastic and even aerodynamic performance without any weight penalty. On the other hand, studies on active control technology have also demonstrated possibilities of significant benefits through control and suppression of aeroelastic instabilities. However, with all the new developments, areas of overlap between structures, aerodynamics and controls in the unsteady, dynamic regime have increased leading to possibilities of aero-servo-elastic (ASE) interactions. Thus an efficient aircraft design calls for an integrated design approach involving close interaction amongst aerodynamic, structural and control engineers at all stages of design. Emergence of new tools such as powerful computers and specialised application software capable of handling a multitude of design variables and parameters have made such an approach possible.

This paper briefly discusses the progress in the last decade in the following two areas: (i) aeroelastic tailoring and structural optimisation with aeroelastic constraints and (ii) aeroservoelasticity. The discussions are restricted to fixed wing aircraft. Recent work and developments in India in the above two areas are described in some detail.

## 2. Aeroelastic tailoring and optimisation

Shirk *et al* (1986) give a standard definition of aeroelastic tailoring as "the embodiment of directional stiffness into an aircraft structural design to control aeroelastic deformation, static or dynamic, in such a fashion as to affect the aerodynamic and structural performance of that aircraft in a beneficial way". Two technological developments in the last two decades have contributed to the evolution of aeroelastic tailoring as a design tool with great potential. These are (i) high performance fibrous composite materials and (ii) mathematical programming methods. The first has significantly increased aircraft structural design options, and the second has permitted the resulting multitude of design variables to be considered and used efficiently. Shirk *et al* (1986) give a comprehensive survey of the status of aeroelastic tailoring, covering the historical background, theoretical foundations and associated research studies on trends, and specific applications of the technology in design exercises. Haftka (1986) also presents developments in the United States in the application of structural optimisation under aeroelastic constraints.

Research studies reported by Shirk *et al* (1986) have shown aeroelastic tailoring using advanced composites to be effective in flutter control, divergence control, particularly of forward swept wings, manoeuvre load relief, improvement of control effectiveness and aerodynamic performance improvement. Figure 1 indicates the aeroelastic deformation trends to be obtained through forward or aft orientation of the primary stiffness direction in a fibre composite laminate with respect to the elastic axis and the corresponding benefits to be expected. Aft location of the primary stiffness direction results in a bend-twist coupling leading to a wash-in (increase of streamwise angle of attack along span) type of torsional deformation under an upward bending load. A wash-out type of torsion (decrease of streamwise angle of attack) is obtained under the same load with a forward location of the primary stiffness direction. It is generally seen that a design with increased aerodynamic effectiveness or improved flutter performance is poorer in terms of manoeuvre load relief and divergence characteristics. Weisshaar & Foist (1982) developed a nondimensional stiffness cross-coupling parameter bounded between  $\pm 1$ , in terms of the bending and torsional stiffness terms and a stiffness cross-coupling term which permitted tailoring studies to be conducted without a detailed definition of the laminate geometry.

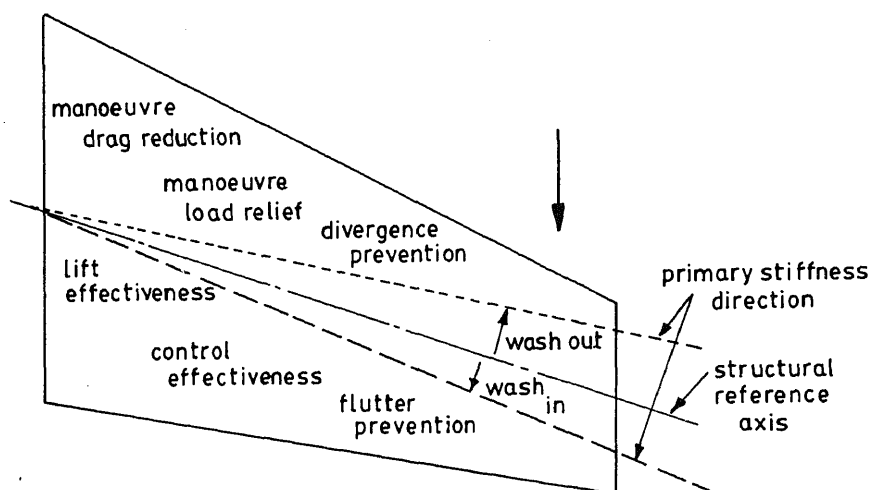


Figure 1. Benefits of aeroelastic tailoring (Shirk *et al* 1986).

## 2.1 Software tools

Use of composite materials, while providing the structural designer with significant design freedom, leads to a significant increase in the complexity of the design problem because of the large number of design variables involved (ply orientations, number of plies, stacking sequence etc). As a result, advanced optimisation techniques have become essential for the efficient design of composite structures. Most of the aircraft companies have developed their own in-house optimisation software to support their design activities. Of these, the following require special mention.

Wing aeroelastic synthesis procedure (TSO) – developed by General Dynamics under an US Airforce contract (McCullers & Lynch 1974), is a preliminary design tool that employs a Ritz equivalent plate model of the wing and nonlinear programming techniques in optimisation. Minimum weight skin thickness and orientations of various plies are calculated subject to a variety of constraints including aeroelastic efficiencies, lift-curve slope and flutter and divergence speeds. The TSO program was used in several design studies for aeroelastic tailoring applications at General Dynamics (YF-16, FB-111), Boeing (KC-135, X-Wing aircraft) and MBB.

FASTOP (flutter and strength optimisation procedure) (Wilkinson *et al* 1977), developed by Grumman, is a finite-element-based two-step design procedure which uses the fully stressed design (FSD) criterion in the first step and a uniform flutter velocity derivative optimality criterion in the second step. The software provides a better structural idealisation as compared to TSO, but the objective function and the constraints are not as varied. FASTOP has been applied by Grumman in several design studies (FSW X-29, Israeli Lavi Fighter) and was also evaluated by Rockwell and Boeing.

ELFINI (Petiau 1990) is a finite-element-based general purpose software developed and extensively used by Dassault Aviation, France, in all its aircraft projects. It presents an integrated approach to aircraft structural analysis and design, incorporating airload computations, aeroelasticity and manoeuvre calculations, flutter analysis and structural optimisation for minimum weight based on a nonlinear projected conjugate gradient method. The design variables are the number of plies

in each fibre direction and the constraints include aeroelastic efficiencies, natural frequencies, flutter speed and damping.

ASTROS (automated structural optimisation system) (Neill *et al* 1990) was developed under US Airforce contract. ASTROS is a multidisciplinary finite-element-based software system that can be used in the preliminary design of aerospace structures. It has primarily integrated existing methodologies into a unified package and, like ELFINI, provides basic static and dynamic structural analysis, steady and unsteady airloads, flutter analysis, sensitivity analysis, optimisation capabilities and aeroservo-elastic response analysis. The optimisation procedure employs a method of modified feasible directions and optimality criteria methods. Optimisation constraints include flutter and divergence speeds, aeroelastic trim parameters and aeroelastic stability derivatives. Ply orientation in composite skin laminates is not a possible design variable as in TSO.

ADOP (aeroelastic design optimisation program) (Dodd *et al* 1990) is an interdisciplinary optimisation program for static, dynamic and aeroelastic analysis using finite element structural models. Developed by the Douglas Aircraft Company, ADOP incorporates FSD concepts, static aeroelasticity and flutter constraints in optimisation. The strategy is to perform a series of optimisation studies rather than attempt the simultaneous satisfaction of multidisciplinary constraints.

ECLIPSE (Kerr & Thomson 1986) is a finite-element-based structural optimisation program developed by the Warton Division of British Aerospace (BAe). The program is optimality criteria based and is capable of dealing with stiffness-based criteria such as aero efficiency, roll rate, divergence speed, flutter speed, natural frequency and frequency separation. The program has been used extensively by BAE in its aircraft design projects.

A finite-element-based software called ASAT (Godel & Schneider 1981) exists at MBB which is able to size cantilevered or free-free surface structures for flutter speed, deflection and strength constraints. CFC structures can also be treated efficiently.

## 2.2 Specific applications

Some specific applications of aeroelastic tailoring and optimisation with aeroelastic requirements in aircraft design during the past 15 to 20 years are briefly described below.

The HiMAT (highly manoeuvrable advanced technology aircraft) remotely piloted research vehicle (Lockenhauer & Layton 1976; Shirk *et al* 1986), designed and built by Rockwell for NASA, was a test-bed for application of aeroelastic tailoring concepts in the lifting surfaces of a modern aircraft. The outboard wing and the canard (figure 2) of HiMAT were aeroelastically tailored to provide optimum transonic manoeuvre performance (load factor of  $8g$  at Mach 0.9 and an altitude of 20,000 ft) with minimum adverse effect on cruise performance. In the design, the aeroelastic twist distribution of the wing was maximised in order to minimise the built-in twist. The TSO program was used in the preliminary design of the vehicle. The required wing ply orientation is shown in figure 2. The HiMAT programme demonstrated the feasibility of unconventional, unbalanced, graphite-epoxy laminates in controlling aeroelastic twist.

Feasibility of designing a forward-swept wing avoiding divergence instability through use of aeroelastic tailoring was demonstrated in the fighter-class aircraft, X-29, developed by Grumman (Hertz *et al* 1982). Grumman used FASTOP in the

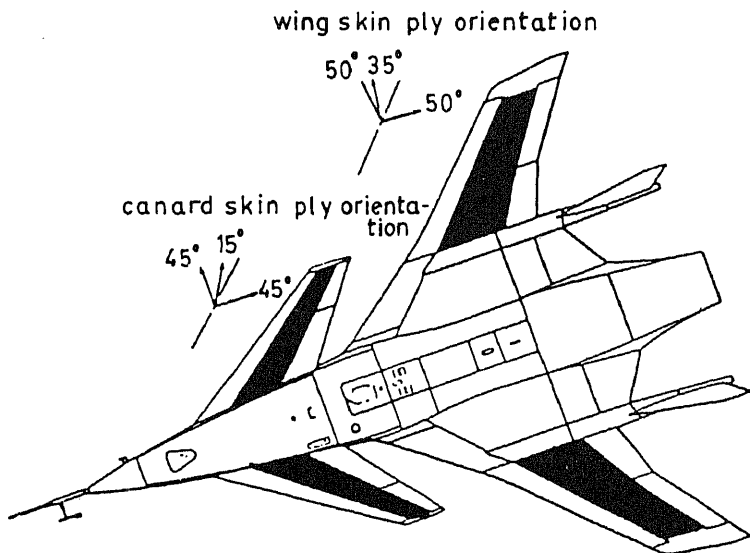


Figure 2. HiMAT remotely piloted research vehicle (Shirk *et al* 1986).

design and arrived at a conventional ( $0^\circ$ ,  $\pm 45^\circ$ ,  $90^\circ$ ) laminate, rotated such that the primary bending plies are  $9^\circ$  forward of the reference structural axis. This provided the necessary bend-twist coupling to minimise the inherent wash-in tendency of the forward-swept wing. A study (Ashley 1982) by Rockwell on a similar concept using TSO also led to a similar result (figure 3).

Applications of advanced optimisation techniques with composites in a combat aircraft wing & fin using ELFINI at Dassault Aviation are presented in Petiau (1990).

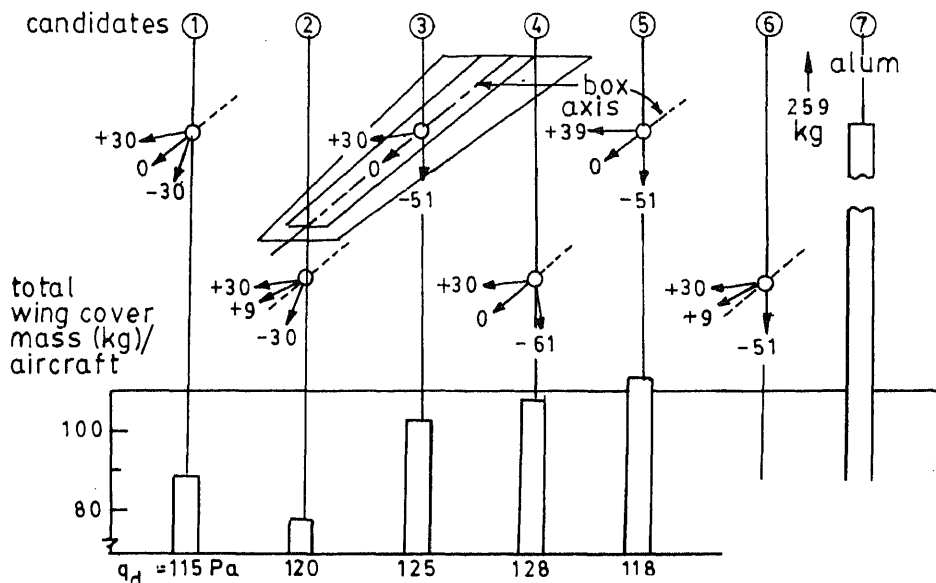


Figure 3. Rockwell FSW fighter wing skin study (dynamic pressure).

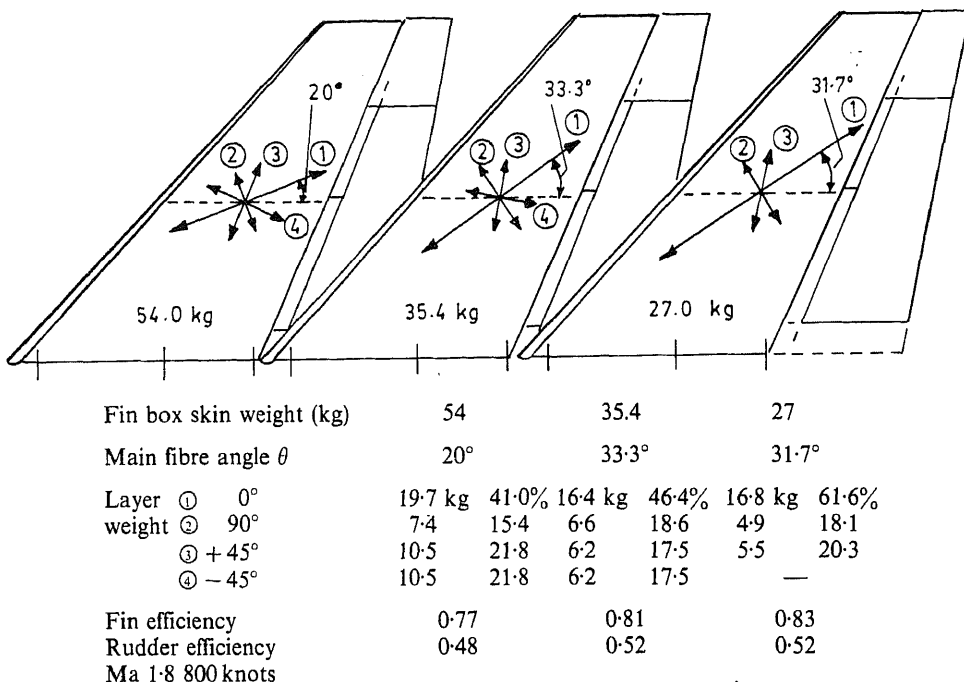


Figure 4. Fin-box skin optimisation using TSO (Schneider & Zimmermann 1986).

The constraints considered include static aeroelasticity, flutter and dynamic frequencies.

Schneider & Zimmerman (1986) describe the use of TSO and FASTOP at MBB in the aeroelastic design studies of the fin and wing boxes of a fighter aircraft. The fin skin optimisation results (figure 4) demonstrated significant benefits from aeroelastic tailoring. Further optimisation for flutter with FASTOP showed an 8.2% increase in flutter speed with only a 2.8% increase in the fin weight.

### 2.3 Developments in India

Efforts in recent years at the Aeronautical Development Agency towards incorporation of aeroelastic tailoring and optimisation concepts in the preliminary design phase of an advanced fighter aircraft are described in this section. The aeroelastic analysis is carried out using the general purpose computer software ELFINI.

The aircraft has a tailless delta-wing configuration with two elevons on each wing for longitudinal and roll control. The fin and rudder provide lateral stability and control. Both the wing and the fin are multi-spar configurations (figure 5) and are of predominantly CFC construction. The skins are tailored and optimised for minimum weight subject to a variety of constraints and load cases.

The aeroelastic analysis uses a finite-element structural model of the aircraft (figure 6) with about 15,000 degrees of freedom in the half symmetric model. Several basic mass distributions are defined on the model which can later be combined to create any defined design mass case.

Basic linear aerodynamic computations based on the method of singularities provide  $\Delta C_p$  distributions on two-dimensional aerodynamic meshes of wing and fin.

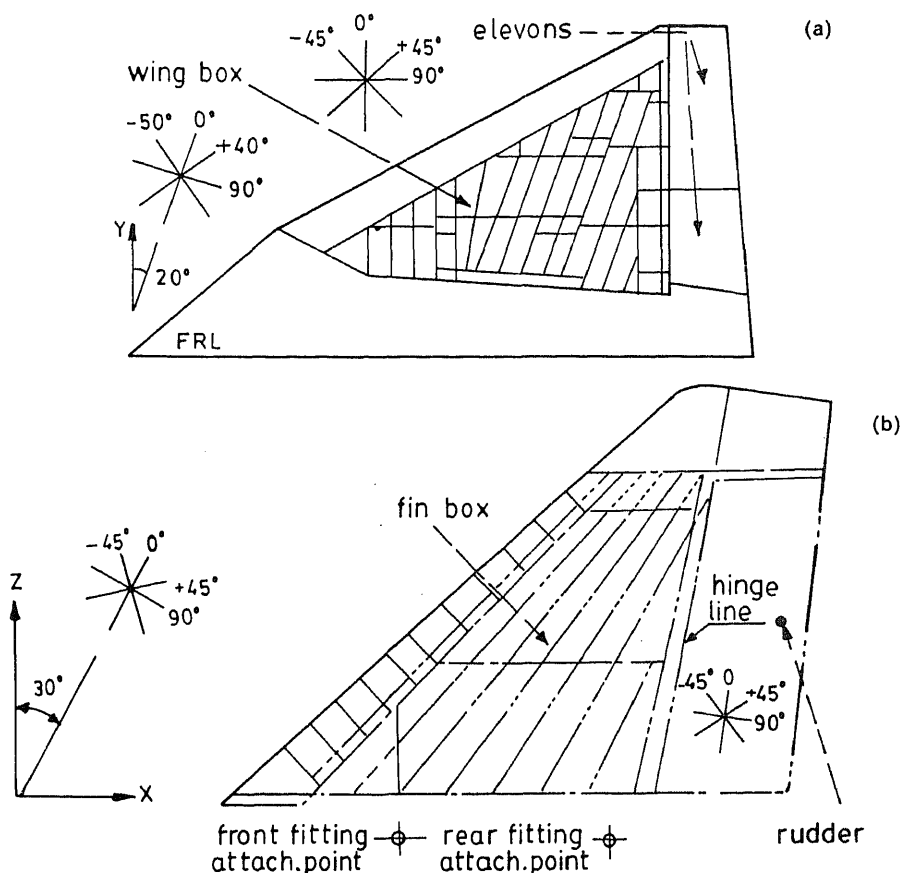


Figure 5. Fighter aircraft wing (a) and fin (b) geometry.

These distributions are corrected (or replaced) using wind-tunnel test data and results from rigorous CFD codes. Bodies such as fuselage, stores etc. are loaded by specifying force and moment resultants on aerodynamic boxes which are subsequently distributed to finite element nodes encompassed by the boxes.

The strategy for aeroelastic analysis involves a two-stage computational procedure; extensive aerodynamic and structural computations are done independently in stage 1 and stored as tables of operators which are subsequently used in stage 2 for less extensive computations of fluid-structure interactions and related aspects for specific mass cases and flight points. This is made possible with the definition of an independent computational grid on which an aerodynamic shape basis and a reduced load basis are defined and which acts as a link between finite element and aerodynamic meshes. Upadhy *et al* (1990a, pp. 331–7) describes in detail the aeroelastic analysis procedure adopted.

Typical results are presented in figures 7 to 9. For an applied rigid  $\Delta C_p$  distribution on the fin (figure 7) for unit values of sideslip  $\beta$  and rudder deflection  $\delta$ , at Mach 1.2, the induced  $\Delta C_p$  distribution due to flexible deformation of the fin and rudder are shown in figure 8. Aeroelastic losses occur mainly due to torsion and chordwise bending of the lifting surface and could be quite large, particularly for control aerodynamic derivatives, at supersonic Mach numbers and high dynamic pressures

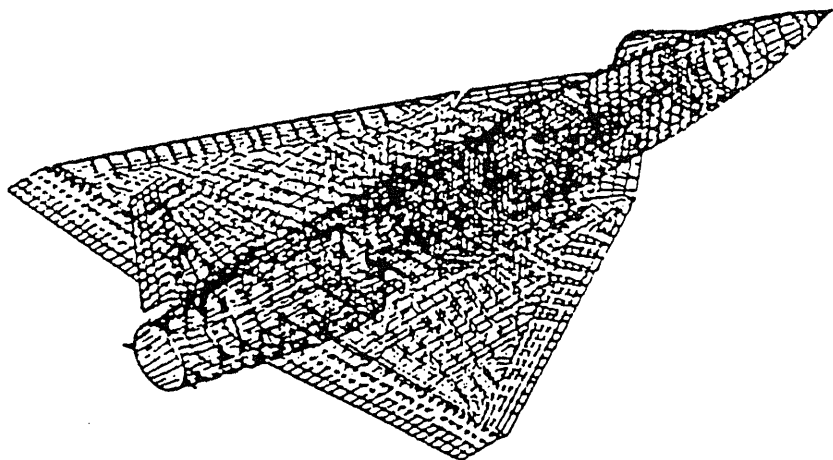
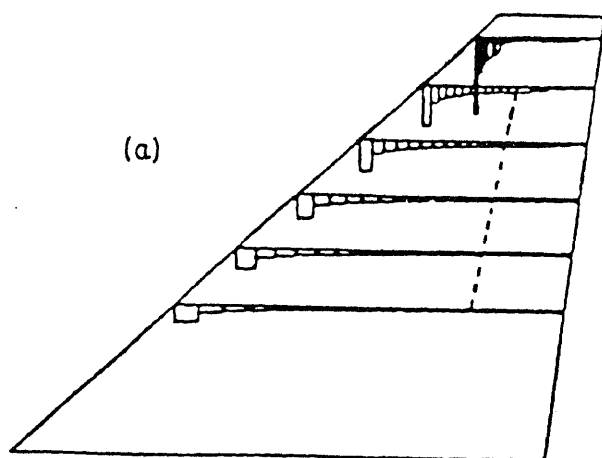


Figure 6. Finite element model.



(a)

(b)

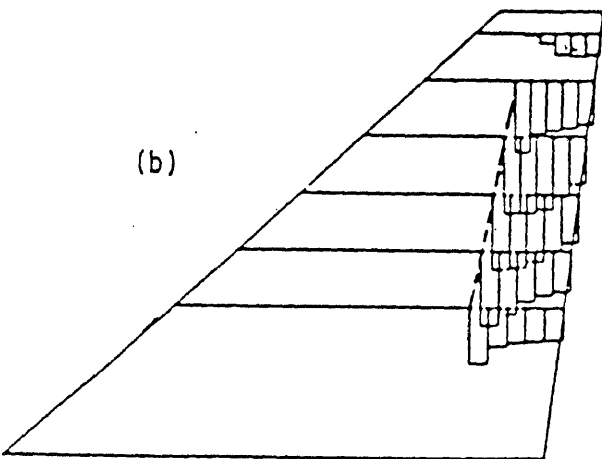
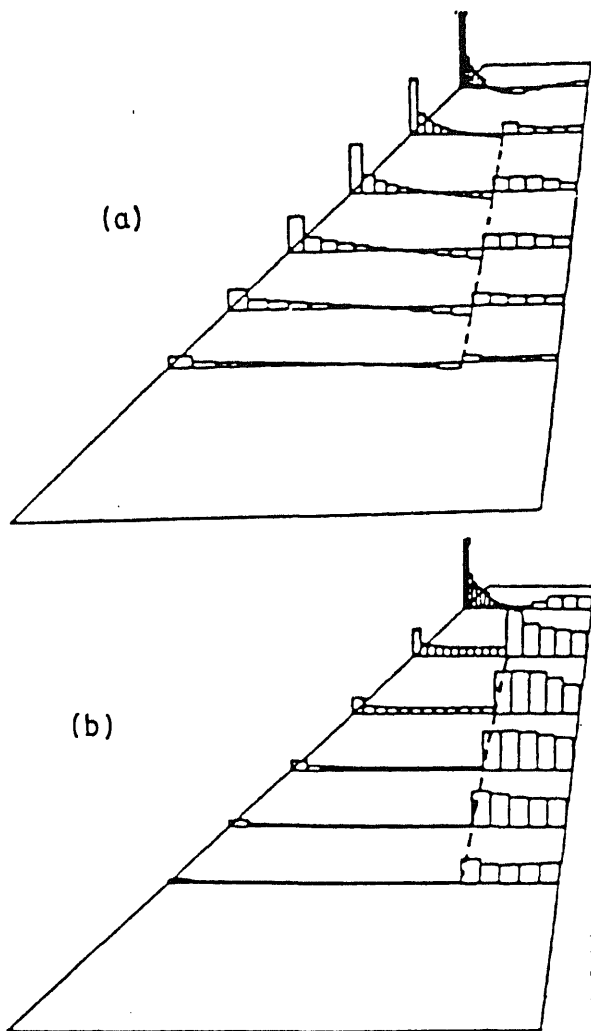


Figure 7. Applied rigid  $\Delta C_p$  distribution on fin ( $M = 1.2$ ). (a) Sideslip  $\beta$ , (b) rudder  $\delta_r$ .



**Figure 8.** Induced  $\Delta C_p$  distribution due to flexibility ( $M = 1.2$ , maximum dynamic pressure). (a) Sideslip  $\beta$ , (b) rudder  $\delta_r$ .

as shown in figure 9. The fly-by-wire FCS of the aircraft will define the minimum acceptable values of aeroelastic efficiency on critical control derivatives and the CFC skins are optimally designed to provide the necessary stiffness and aeroelastic deformation coupling. Typical aeroelastic deformation shapes of the aircraft for angle of attack ( $\alpha$ ) and antisymmetric elevon ( $\delta_a$ ) effects are presented in figure 10.

The optimisation procedure requires definition of zones or patches on the skins (figure 11). The design variables are the thicknesses (or number of plies) in each of the four pre-defined fibre directions in the zones. The total number of design variables are 328 for the CFC wing skins and 154 for CFC fin skins.

Critical load cases involving combinations of normal load factor and roll rate/acceleration for the wing and sideslip and rudder deflection for the fin are defined for the optimisation study.

Static aeroelastic requirements in optimisation are defined as minimum acceptable values of elevon effectiveness in roll or rudder effectiveness in yaw. These parameters design the torsional stiffness of wing/fin box and also the bend-twist coupling which



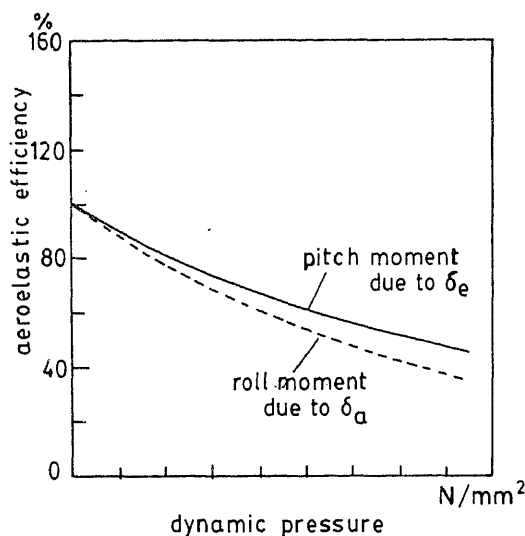


Figure 9. Aeroelastic efficiency variation.

determines the wash-in/wash-out characteristics. Other constraints considered in optimisation are failure criteria in the skins (modified Tsai–Hill criterion for composite plies), local buckling of skin panels between spars and limited technological constraints such as maximum and minimum thickness in any given direction as a ratio of total laminate thickness in a given zone. Aeroelastic constraint is an overall constraint on the structure and is independent of load cases. Details of the optimisation procedure are presented in Upadhyaya *et al* (1990b, pp. 485–90).

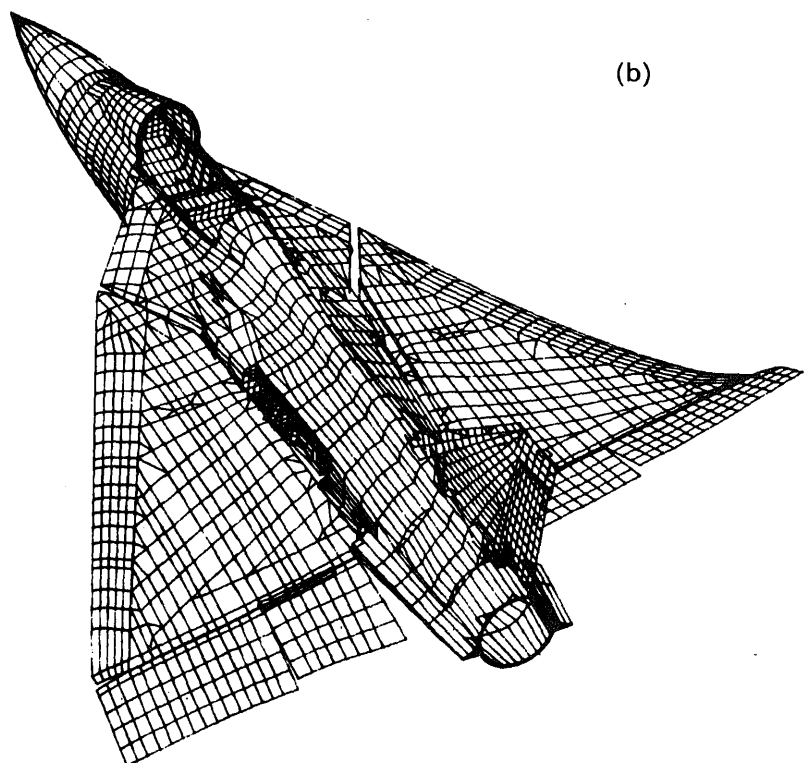
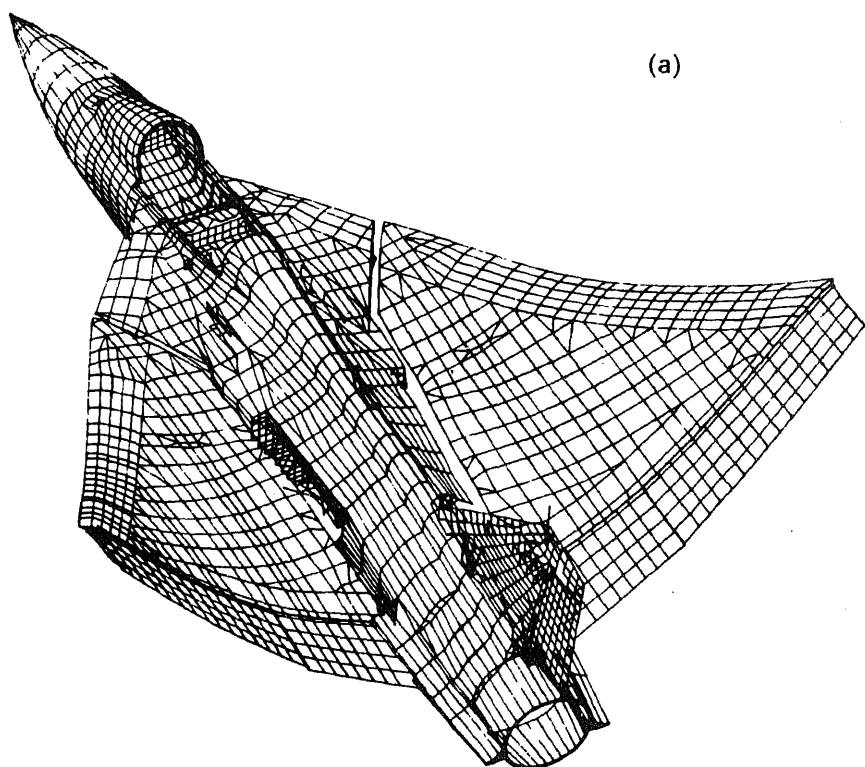
The results of various aeroelastic tailoring/optimisation studies on the wing are summarised below.

(1) Table 1 presents evolution of optimum weight with design constraints for a generally used value of aeroelastic efficiency parameter. These results pertain to a  $0^\circ/+40^\circ/-50^\circ/90^\circ$  lay-up.

(2) Figure 12 compares optimum weights of metallic (Al–Cu and Al–Li) and CFC (T300 & T800) wing skins taking Al–Cu skin weight for roll control efficiency value C as reference. Value B corresponds to table 1 and a 5% variation on either side is considered.

It is seen that (i) for a given material, weight increase is steeper at higher values of roll efficiency as more zones come under the influence of this constraint (figure 13), (ii) the increase in weight with roll efficiency is much smaller for composites when compared to metals. This arises from the advantage of aeroelastic tailoring with composites where only the thickness in a required direction is increased to meet aeroelastic requirements.

(3) Table 2 compares relative weights of wing skins for various types of lay-ups. It is seen that unbalanced  $0^\circ/+45^\circ/-45^\circ/90^\circ$  lay-up gave the minimum weight. Balanced lay-up resulted in a 31% weight penalty. Comparison of weights of lay-ups 1, la, 2 and 3 showed that orienting the  $0^\circ$  fibres along the inclined spars is the optimum arrangement. The results also showed that in lay-up 1, the  $+45^\circ$  direction was the most predominant one, accounting for 43.6% of the weight, followed by  $-45^\circ$  (28.2%),  $90^\circ$  (15.6%) and  $0^\circ$  (12.6%) layers. In the case of balanced lay-up, both  $\pm 45^\circ$  layers shared 37.2% of the weight. It is thus clear that while both  $\pm 45^\circ$  layers contribute towards torsional stiffness, the predominant  $+45^\circ$  layers in lay-up 1 also provide



**Figure 10.** Aeroelastic deformation shapes,  $\alpha$ (a) and  $\delta_a$  (b) effects.

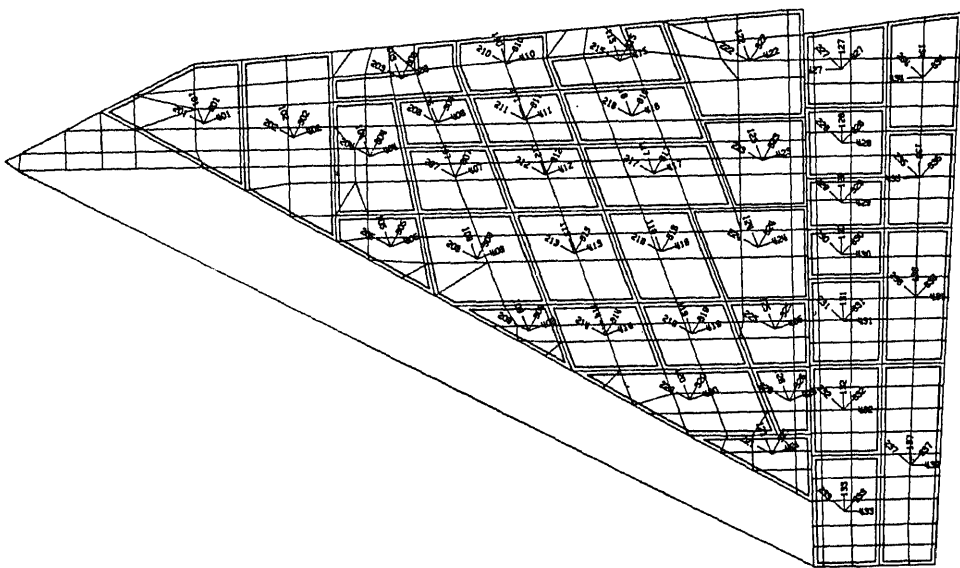


Figure 11. Optimization zones on wing skins.

the required twist-bend coupling, causing a reduced nose-down twist (i.e., effective reduction in angle of attack) under an up-load on the elevon, leading to a lower weight structure for a given elevon effectiveness. In other lay-ups also, it was seen that the  $+ \theta$  layers were the most predominant.

(4) The wing tip displacements in a critical load case for design case B for the Al-Cu, Al-Li, T300 and T800 material designs were 179, 176, 210 and 201 mm, respectively. The higher wing tip deflections in the composite wing are again due to the tailoring effect where the bending stiffness is not increased in the same proportion as the torsional stiffness like in metal wings.

Aeroelastic analysis and optimisation studies on the fin (Kamal *et al* 1989) showed trends similar to that seen on the wing with the  $+45^\circ$  layers being predominant. This design results in a coupled bending of the fin due to torsional load from the rudder, which induces a streamwise incidence opposed to the twist due to torsion, thus reducing the losses in the yaw moment effectiveness of the rudder due to flexibility. Another interesting observation was that rotating the entire lay-up, such that the

Table 1. Evolution of optimum weight. Lay-up:  $0^\circ/+40^\circ/-50^\circ/90^\circ$  on wing-box skin.

No.	Design constraints	Weight (ratio)
1.	Composite material – stress (failure criteria constraints only)	0.46
2.	Stress + buckling constraints	1.0
3.	Stress + buckling + aeroelasticity	1.25
4.	3 + Technological constraints	1.31
5.	Aeroelastic constraint only	0.90
6.	Aluminium alloy skin (comparable with 4)	2.25

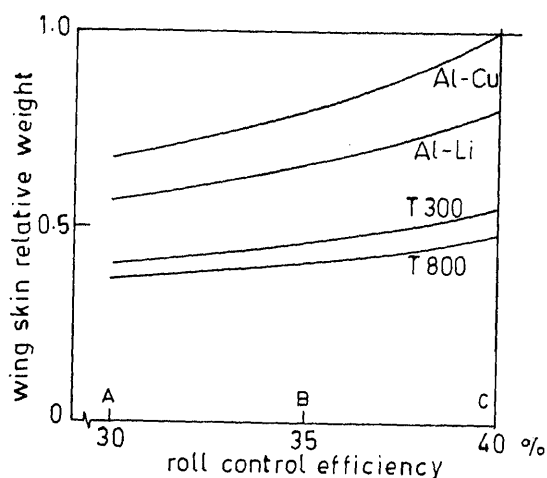


Figure 12. Comparison of optimum design wing skin weights.

angle between  $0^\circ$  fibres and the Z-axis varied in the range  $15^\circ$  to  $35^\circ$ , showed minimum skin weight (9% weight saving as compared to the  $30^\circ$  case) for the  $15^\circ$  orientation. This is most likely due to the fact that the predominant  $+45^\circ$  layers provide relatively more direct bending stiffness for this orientation, resulting in less induced spanwise incidence due to direct bending of the elastic axis due to sideload on the rudder. The most optimum fibre orientation is one which is a compromise between the coupled bending and direct bending effects which induce spanwise incidence in opposite directions.

The study also revealed that the fin skin weight is very sensitive to the aeroelastic

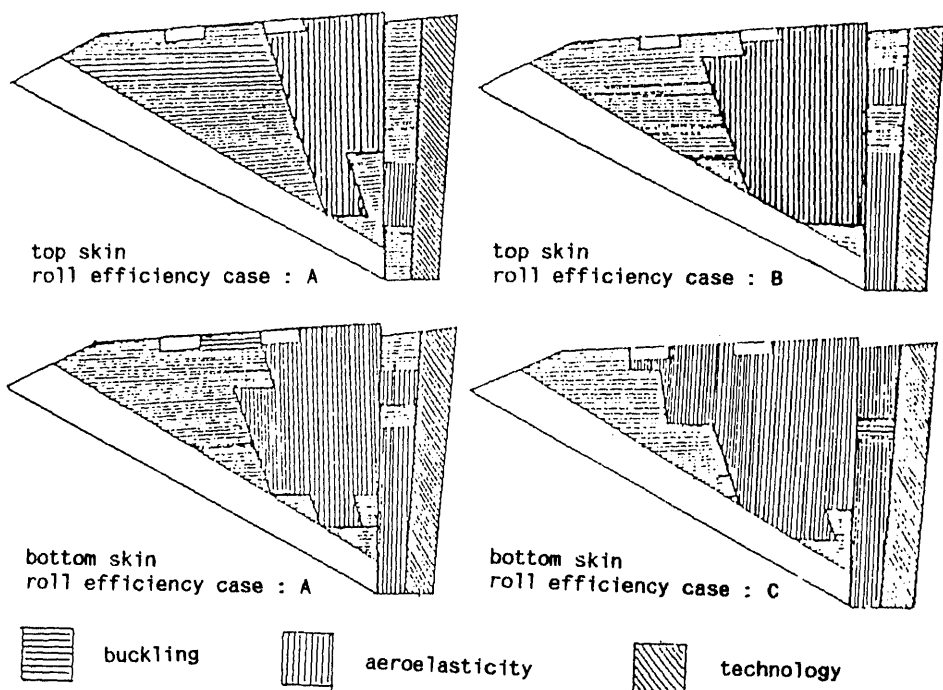


Figure 13. Optimisation constraint influence zones.

**Table 2.** Relative weight comparison.

Wing-box skin lay-up	Relative weight
$0^\circ/+45^\circ/-45^\circ/90^\circ$	1.00
$0^\circ/+45^\circ/-45^\circ/90^\circ$ (balanced)	1.31
$10^\circ/+45^\circ/-45^\circ/90^\circ$	1.20
$-10^\circ/+45^\circ/-45^\circ/90^\circ$	1.25
$0^\circ/+40^\circ/-40^\circ/90^\circ$	1.23
$0^\circ/+40^\circ/-50^\circ/90^\circ$	1.18
$0^\circ/+50^\circ/-40^\circ/90^\circ$	1.29
$0^\circ/+50^\circ/-50^\circ/90^\circ$	1.25
$0^\circ/+22.5^\circ/-67.5^\circ/90^\circ$	1.65

Note: Reference axis for lay-up is inclined  $20^\circ$  aft of Y-axis (figure 5).

efficiency parameter. A requirement of 5% increase in supersonic yaw moment effectiveness of rudder resulted in a fin skin weight increase of nearly 36%.

### 3. Aeroservoelasticity

For the modern aircraft, use of an efficient flexible structure is possible with the availability of composite materials and modern design and analysis techniques, aided by large capacity, high speed computers and sophisticated structural testing methods. As the structure became more flexible, the structural frequencies have come down. At the same time advances in aircraft design have necessitated expansion of the role played by flight control systems (FCS). The aircraft is designed to be highly manoeuvrable, light, and more agile. For such an aircraft the flight control system is not just used for the control of the flight path, but also to meet the above design requirement. Thus the role of FCS has been expanded. High authority control systems, utilizing multiple blended feedbacks, are used to provide tailored aircraft response to meet mission requirements. In addition, the flight control system is used to provide artificial stability for statically unstable aircraft, reduce fatigue damage to the structure from gusts and manoeuvres, suppress flutter modes and improve ride quality. This active control technology helps to reduce structural weight of the aircraft, and provides a better aerodynamic design. With the introduction of high gain, faster response flight control systems to meet the above active control requirements, the control frequency bandwidth has increased. Further, as the frequency of operation of the control system increases, it excites the aircraft structure at higher frequencies, as a result of which the reduced frequency characterising the flow unsteadiness becomes significantly large indicating that unsteady aerodynamic effects need to be accounted for. The reduction in structural frequencies and coupled increase in the control system frequency bandwidth with the presence of unsteady aerodynamics leads to a new dynamic interaction problem which is defined as aeroservoelasticity (ASE). Thus, ASE is a multidisciplinary technology dealing with the interaction of the aircraft's flexible structure and the unsteady aerodynamic forces resulting from the motion of the aircraft with its flight control system. Detailed and complex mathematical models incorporating the effects of these technical disciplines are required to accurately predict ASE interactions and to design active control systems for flexible vehicle application.

### 3.1 ASE encounters

Several instances of ASE encounters for a variety of research, development, prototype and production aircraft are documented in literature. Felt *et al* (1979) have discussed in depth ASE encounters for the following aircraft:

**B-36:** An ASE instability was induced by the autopilot of B-36. For this aircraft the sensor package had been located in the tail gunner's compartment and significant body bending (flexible) motion had been picked up by the sensor. The solution was to move the sensor package to a position of relatively small body bending motion.

**YF-16:** This aircraft experienced two separate ASE instabilities during early flight tests. Initially these instabilities were not identified by analysis, because the analysis was carried out for high Mach number, low altitude flight conditions which were the most critical for flutter. However, a critical interaction occurred at a high subsonic Mach number, which was not considered in analysis. Later ASE analysis at the proper flight condition predicted the ASE instabilities which matched with the flight test results.

**YF-17:** YF-17 experienced two ASE instabilities, which were predicted during the analysis stage and verified with flight test results. Parametric ASE analysis of ground test and flight test interactions were carried out to select final notch filters for the flight control system.

**B-52:** The CCV ride quality system tested on a B-52 encountered servoelastic oscillations on ground due to local structural vibration of bulkheads or support beams. As the local structural details were not included in the dynamic mathematical model, this interaction was not identified by the analysis. By changing accelerometer mounting, the oscillations were eliminated.

**F-4:** ASE instability was encountered in modified F-4 aircraft during a sideslip manoeuvre in a gear down, flap down configuration. A resonance in the pitch axis was encountered. The instability occurred at 23 Hz, which was close to both the flap rotation mode and the stabilizer rotation mode frequencies. The instability mechanism was initiated by a flap buffet which fed into a pitch rate gyro located in the left wing root just forward of the flap. The problem was solved by additional filtering in the pitch axis.

**F-16 and F-18:** These aircraft also exhibited adverse dynamic interactions between the airframe aeroelastic characteristics and their flight control systems. Flight control modifications were carried out to solve this interaction problem.

### 3.2 General formulation of ASE problem

Very few references are available in literature on ASE formulation. Noll *et al* (1989) present activities in the area of ASE at the NASA Langley Research Centre. They have discussed different modelling and analysis techniques for the prediction of ASE interactions. Suryanarayan *et al* (1992) have discussed in detail various analysis steps required to be carried out for ASE analysis. The steps are given starting from aerodynamic and control system modelling up to identification of

the flight envelope. The steps are as below.

- (i) Normal mode analysis of the aircraft structure for a given mass case (mode shapes, frequencies, modal mass and modal stiffness computation).
- (ii) Calculation of modal unsteady airloads as a function of reduced frequency and Mach number for harmonic motion.
- (iii) Representation of modal unsteady airloads in the time domain for arbitrary motion.
- (iv) Dynamic modelling of the actuator assemblies.
- (v) Development of a flexible aircraft dynamic model taking into account the effects of elastic modes and unsteady airloads along with the actuator model (state-space model of flexible aircraft).
- (vi) Open loop poles and zeros computation, formation of open loop transfer function for the flexible aircraft.
- (vii) Calculation of open loop frequency response at the sensor location for a sinusoidal input to the actuator from the state-space model as well as from the open loop transfer functions.
- (viii) Representation of control system dynamics in terms of transfer functions.
- (ix) Development of closed loop aircraft description taking all the feedback element dynamics into account.
- (x) ASE stability analysis for various flight points in the flight envelope.

The flexible aircraft state-space model as given in step (v) is used for active control design applications.

### 3.3 ASE analysis tools

The following general purpose analysis programs are documented in the literature for carrying out ASE analysis.

STARS (Gupta *et al* 1989): Structures, aerodynamics and related aeroservoelastic system analysis is a general-purpose structural analysis program that has a complete aeroservoelastic analysis capability. In the package, Pade and least squares approximation methods are available for time domain representation of unsteady aerodynamics and it is used for open and closed loop aeroservoelastic controls analysis.

ADAM and ADAM-2.0 (Noll *et al* 1986) (Analog and digital aeroservoelasticity method): This program combines the technologies of unsteady aerodynamics, multi-input/multi output (MIMO) controls and structural dynamics into an interactive analysis package.

ISAC (Interaction of structures, aerodynamics and control) (Sallee 1990): All the active control problems at NASA Langley are analysed using ISAC. The package has 3 methods for performing aerodynamic approximation and uses optimal method for control law synthesis.

ASEPACK, Aeroservoelasticity package (Suryanarayan *et al* 1992): This program has been developed at the Aeronautical Development Agency. The program takes structural dynamics and frequency domain unsteady aerodynamics input from the general purpose finite element package ELFINI. The package has been developed on similar lines to STARS. Various in-house developed new analysis techniques, namely, modified rational function approximation (RFA) for time-domain representation of unsteady aerodynamics and low-order aircraft plant model design for the use of flight

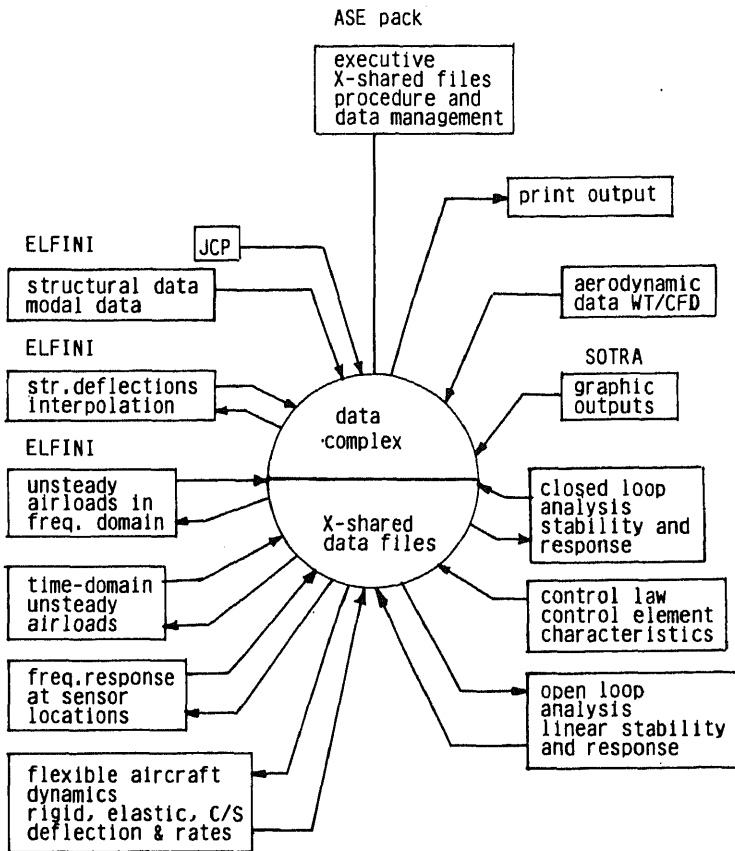


Figure 14. ASEPACK block diagram (Suryanarayan *et al* 1992).

Table 3. Comparison of ASEPACK and STARS programs.

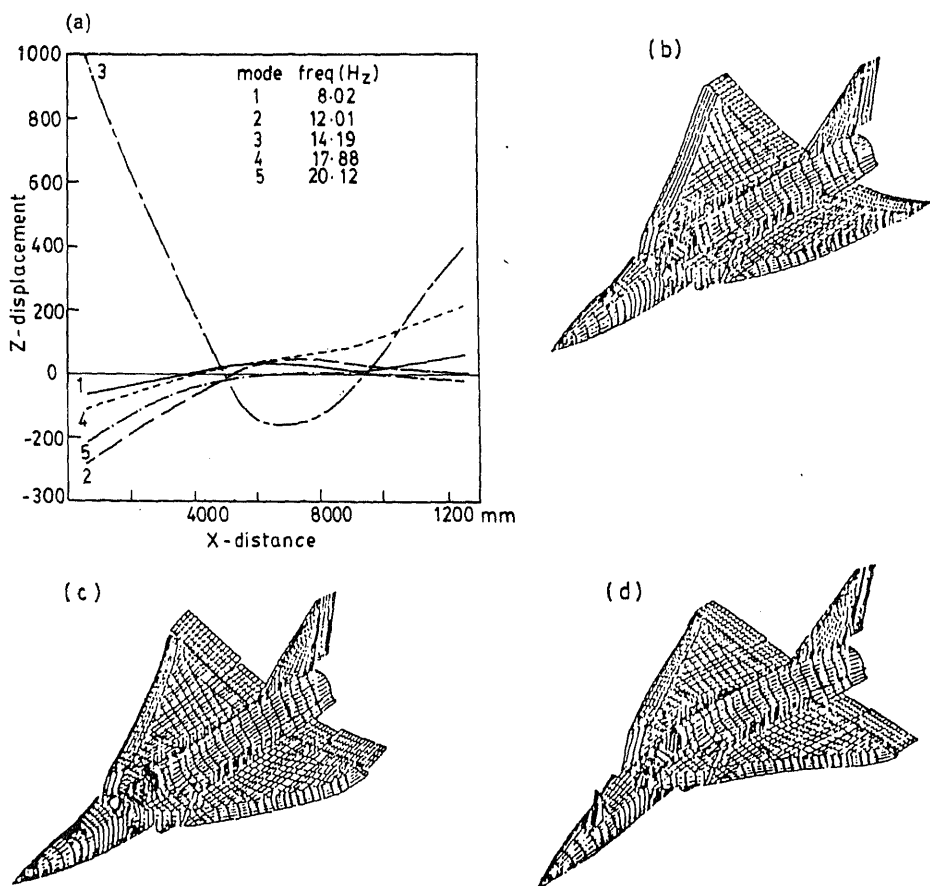
Capability	ASEPACK	STARS
O/L and C/L S-plane stability analysis	Yes	Yes
Loop closure with feedback elements	Yes	Yes
Bode plots of O/L and C/L frequency response	Yes	Yes
Maximum state variables	No limit	100
Aerodynamic lag terms	No limit	4
Time history analysis	Yes	Yes
Actuator representation	Fourth order	Fourth order
Sensor dynamics	Yes	Yes
Anti-aliasing filter	Second order	First order
Notch/shaping filter	Yes	Yes
Subsonic/supersonic	Yes	Yes
O/L and C/L Z-plane stability analysis	Yes	Yes
Time domain representation	Pure lag	Conventional
Reduced order plant representation	Yes	No



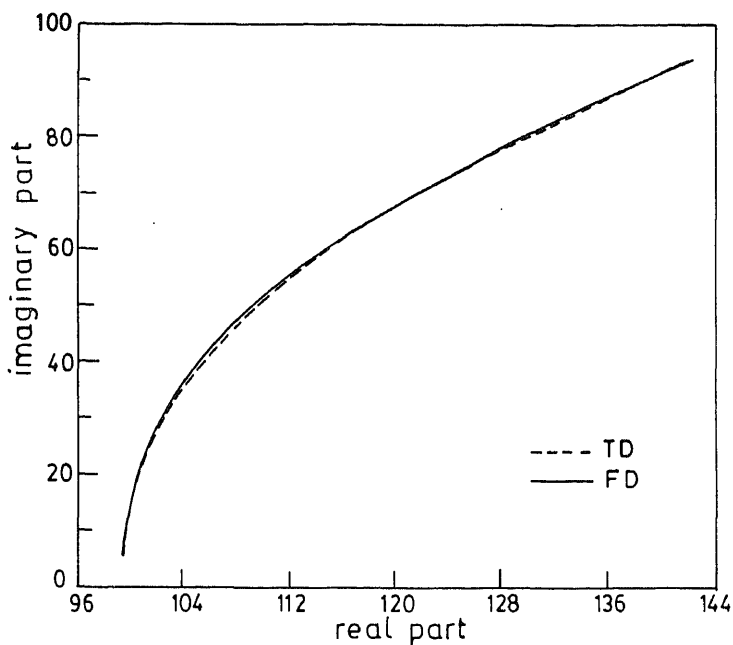
control-law design and simulation studies are implemented in this package. Figure 14 shows the ASEPACK block diagram. Table 3 shows a comparison of the capabilities of the ASEPACK and STARS programs.

### 3.4 Application of ASEPACK to a typical fighter aircraft

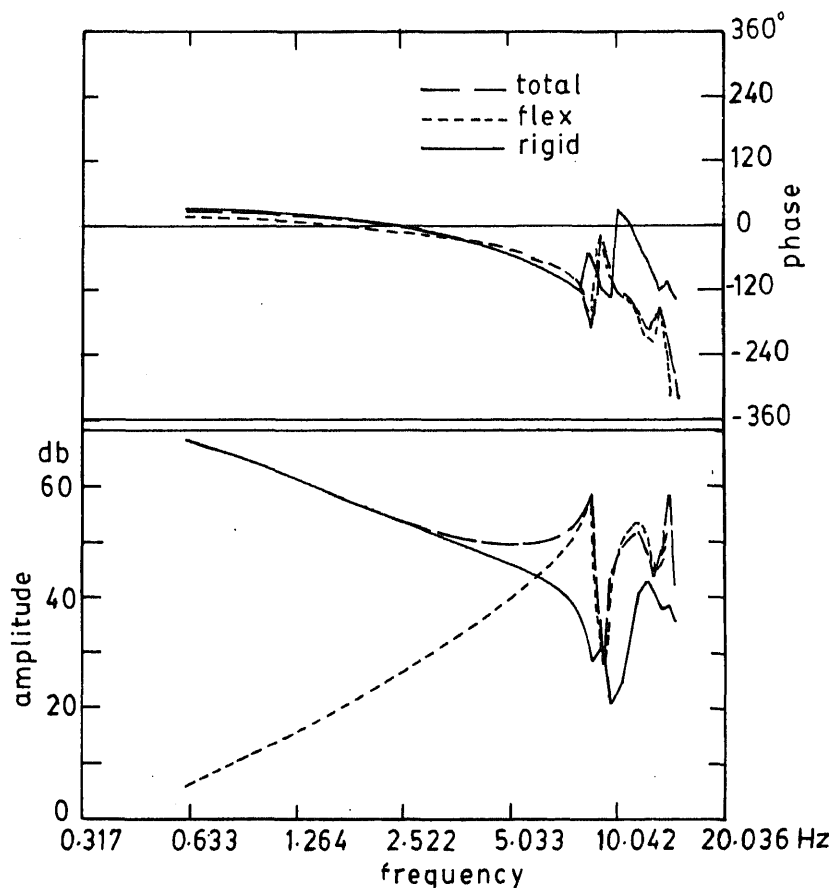
The ASE analysis package is used for analysis of a typical fly-by-wire fighter aircraft configuration. Figure 15 shows the first wing bending and torsion modes, first fuselage bending mode, and also a line representation of fuselage deformation in the first five symmetric modes. Figure 16 shows the accuracy of the time-domain representation of a typical aerodynamic coefficient and figure 17 shows the open-loop normal acceleration (both amplitude and phase) response of the flexible aircraft plant at the sensor location with the rigid and flexible contributions and the total response shown separately. For the present analysis two rigid body modes (heave and pitch) and the first four symmetric elastic modes covering wing bending and twisting, and fuselage bending are considered. Both open-loop and closed-loop stability and frequency response results are generated for a typical flight condition.



**Figure 15.** Free vibration mode shapes (Suryanarayan *et al* 1992). (a) Fuselage symmetric mode shapes; (b) wing bending; (c) wing torsion; (d) fuselage bending.



**Figure 16.** Frequency domain (FD) and time domain (TD) representation of  $C_{L\alpha}$  (Suryanarayan *et al* 1992).



**Figure 17.** Open loop frequency response results (Suryanarayan *et al* 1992).

#### 4. Conclusions

In the preceding sections, recent progress in the areas of aeroelastic tailoring, structural optimisation with aeroelastic requirements and aeroservoelasticity are discussed. Developments and applications of the above technologies in India are highlighted.

So far aeroelastic tailoring and active control technology have progressed independently. However, the aeroelastic benefits to be derived from the two techniques have their own limits. Effective integration of structural stiffness tailoring and active controls through simultaneous design of a structure and control law holds great promise (Livne *et al* 1990; Dracopoulos & Oz 1992). It can be expected that they will merge into a single design tool within the next decade.

The authors are grateful to Dr Kota Harinarayana for encouragement.

#### References

- Ashley H 1982 On making things the best – Aeronautical uses of optimisation. *J. Aircr.* 19: 5–28
- Dodd A J, Kadinka K E, Loikkanen M J, Rommel B A, Sikes G D, Strong R C, Tzong T J 1990 Aeroelastic design optimisation program. *J. Aircr.* 27: 1028–1036
- Dracopoulos T N, Oz H 1992 Integrated aeroelastic control optimisation of laminated composite lifting surfaces. *J. Aircr.* 29: 280–288
- Felt L R, Huttzell L J, Noll T E, Cooley D E 1979 Aeroservoelastic encounters. *J. Aircr.* 16: 477–483
- Godel H, Schneider G 1981 Application of a structural optimisation procedure for advanced wings. AIAA-R-691
- Gupta K K, Brenner M J, Voelker L S 1989 Integrated aeroservoelastic analysis capability with X-29A comparisons. *J. Aircr.* 26: 84–90
- Haftka R T 1986 Structural optimisation with aeroelastic constraints: A survey of US applications. *Int. J. Vehicle Design* 7: 381–392
- Hertz T J, Shirk M H, Ricketts R H, Weisshaar T A 1982 On the track of practical forward swept wings. *Astronaut. Aeronaut.* 20: 40–52
- Kamal K, Srinivasan K, Panda K 1989 Preliminary optimisation study of a fighter aircraft fin. ADA/060200/040/89
- Kerr R I, Thompson D 1986 Automated structural optimisation at Warton. *Proc. 15th Congress of ICAS* (eds) P Santini, R Staufienbiel Paper ICAS-86-3.1.3 (New York: AIAA) 1: 130–134
- Livne E, Schmit L A, Friedmann P P 1990 Towards integrated multidisciplinary synthesis of actively controlled fibre composite wings. *J. Aircr.* 27: 979–992
- Lockenhauer J L, Layton G P 1976 RPRV research focus on HiMAT. *Astronaut. Aeronaut.* 14: 36–41
- McCullers L A, Lynch R W 1974 Dynamic characterisation of advanced filamentary composite structures. Volume II. Aeroelastic synthesis procedure development. AFFDL-TR-73-111
- Neill D J, Johnson E H, Camfield R 1990 ASTROS—A multidisciplinary automated structural design tool. *J. Aircr.* 27: 1021–1027
- Noll T, Blair M, Cerra J 1986 ADAM, An aeroservoelastic method for analog or digital systems. *J. Aircr.* 23: 852–858
- Noll T E, Perry B, Gilbert M G 1989 Aeroservoelasticity: Current trends and future expectations. *Australian Aeronautical Conference*, Melbourne, pp. 176–180
- Petiau C 1990 Structural optimisation of aircrafts—practice and trends. *Proceedings of the 17th Congress of the Int. Council of the Aeronautical Sciences*, Stockholm, Sweden, Paper No. ICAS-90-1.2.1 (New York: AIAA) 1: 210–221

- Sallee V J 1990 ADAM-2.0 – An ASE analysis code for aircraft with digital flight control systems. AIAA-90-1077-CP
- Schneider G, Zimmermann H 1986 Static aeroelastic effects on high-performance aircraft. AGARD-R-725, pp. 2.1–2.15
- Shirk M H, Hertz T J, Weisshaar T A 1986 Aeroelastic tailoring—theory, practice and promise. *J. Aircr.* 23: 6–18
- Suryanarayan S, Panda K, Prakash B G, Sujata T 1992 Aeroservoelastic stability and response analysis: Integration of component analysis and software development. Report No. ADA/TD/STR/003, Aeronautical Development Agency, Bangalore
- Upadhyaya A R, Ramani T S, Srinivasan K 1990a Approach to static aeroelastic analysis of a combat aircraft. *Proc. Int. Conf. on Advances in Structural Testing, Analysis and Design* (New Delhi: Tata-McGraw Hill)
- Upadhyaya A R, Ramani T S, Srinivasan K, Rajasekaran P 1990b Studies on optimisation of an advanced fighter aircraft wing structure. *Int. Conf. on Advances in Structural Testing, Analysis and Design* (New Delhi: Tata-McGraw Hill)
- Weisshaar T A, Foist B L 1982 Vibration and flutter of advanced composite lifting surfaces. AIAA paper 82-0722
- Wilkinson K, Markowitz J, Lerner E, George D, Batill S M 1977 FASTOP: A flutter and strength optimisation program for lifting surface structures. *J. Aircr.* 14: 581–587



## Forced vibration and low-velocity impact of laminated composite plates<sup>†</sup>

ASGHAR NOSIER<sup>†</sup>, RAKESH K KAPANIA<sup>‡</sup> and J N REDDY\*

Aerospace & Ocean Engineering Department, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

\* Present address: Oscar S Wyatt Chair, Department of Mechanical Engineering, Texas A & M University, College Station, TX 77843, USA

**Abstract.** The layerwise theory of Reddy is used to study the low velocity impact response of laminated plates. Forced-vibration analysis is developed by the modal superposition technique. Six different models are introduced for representation of the impact pressure distribution. The first five models, in which the contact area is assumed to be known, result in a nonlinear integral equation similar to the one obtained by Timoshenko in 1913. The resulting nonlinear integral equation is discretised using a time-element scheme. Two different interpolation functions, namely: (i) Lagrangian and (ii) Hermite, are used to express the impact force. The Hermitian polynomial-based representation, obviously more sophisticated, is introduced to verify the Lagrangian-based representation. Due to its modular nature the present numerical technique is preferable to the existing numerical methods in the literature. The final loading model, in which the time dependence of the contact area is taken into account according to the Hertzian contact law, resulted in a relatively more complicated but more realistic, nonlinear integral equation. The analytical developments concerning this model are all new and are reported for the first time in this paper. Also a simple, but accurate, numerical technique is developed for solving our new nonlinear integral equation which results in the time-history of the impact force. Our numerical results are first tested with a series of existing example problems. Then a detailed study concerning all the response quantities, including the in-plane and interlaminar stresses, is carried out for symmetric and antisymmetric cross-ply laminates and important conclusions are reached concerning the usefulness and accuracy of the various plate theories.

**Keywords.** Low-velocity impact; laminated composites; layerwise theory; forced vibration; nonlinear integral equation.

---

<sup>†</sup> This paper, presented at a symposium to celebrate the Golden Jubilee of the Aerospace Department, Indian Institute of Science, Bangalore first appeared as a paper in *Computational Mechanics* (1994) 13: 360–379 © Springer-Verlag. Reproduced here with slight alterations with permission.

## 1. Introduction

Composites are often used in situations involving the sudden application of loads. The dynamic response of the structure ensues after load application and a state of stress leading to failure may be generated. It is necessary to understand the response characteristics of the material body for all important effects, including geometry, boundary conditions, and loading.

One of the major obstacles that prevents application of these materials in primary structures is the damage induced due to service or accidental loads (e.g., bird impact), and the consequent reduction in stiffness, strength and life of these structures. Therefore, damage-resistant and durable composite materials are essential for the design of lighter and easier-to-maintain structures. The composite laminate is a fundamental building block of a composite structure. Hence, an understanding of damage development and failure behaviour in a composite laminate is a basic requirement for understanding the failure behaviour of a composite structure. Investigations in the area of wave propagation in laminated media have been conducted by geologists and physicists interested in the study of the wave propagation of seismic waves. The increasing use of laminated composites in aerospace, automotive and naval structures has led to a more elaborate area of research. These structures are subject to high velocity impact by birds, meteoroids, and undersea animals. Detailed reviews of some of these studies were given by Kapania & Raciti (1989) and Abrate (1991). Only a brief review of some of the pertinent recent analytical studies is presented in the following pages.

The transient response of laminated plates subjected to impact loads was investigated by analytical and numerical methods. Goldsmith (1960) used the normal modes method to determine the dynamic response of an isotropic plate or beam to a rigid impactor. Timoshenko (1913) used normal modes and a Hertzian contact law to analyse the deflections of a beam due to impact. The resulting nonlinear integral equations were solved by numerical integration. Sun & Chattopadhyay (1975) extended Timoshenko's method to a laminated simply-supported composite plate. Ramkumar & Chen (1983) used the Fourier integral transform to find the response of an infinite anisotropic laminated plate to an experimentally determined impact force. Petersen (1985) used the finite method based on a shear deformable plate theory with rotary inertia to analyse laminated plates subjected to impact loads. Thangjiatham *et al* (1987) obtained low-velocity impact responses of orthotropic plates using a higher-order theory that incorporates: (i) the transverse normal stress, (ii) rotary inertia effects, and (iii) fulfills the shear stress boundary conditions on the bounding surfaces. Sun & Liou (1989) used a three-dimensional hybrid stress finite element method to investigate laminated plates under impact loads. Cairns & Lagace (1989) obtained transient response of graphite/epoxy and kevlar/epoxy laminates subjected to impact using the Rayleigh-Ritz method.

For the study of impact response of metals and composites, many researchers used the Hertzian contact law, which relates impactor and plate motion with contact force. However, Yang & Sun (1981) showed that the Hertzian contact law was not adequate by performing statical indentation tests on graphite/epoxy composite laminates using spherical steel indentors of different sizes. They found that significant permanent indentations existed. In order to account for the permanent indentation, Tan & Sun (1982) proposed a modified Hertzian contact law following Yang & Sun (1981). They compared experimental results with the predictions of finite element analysis using the statically determined contact law. Sun & Chen (1985) analysed initially stressed

composite plates under impact loads using this modified Hertzian contact law. Bogdanovich & Yarve (1989) proposed a method which combined the calculation of stress-strain states in a laminated plate on the basis of spline-approximation of displacements. A variational approach was used for studying the process of impact contact interaction between the indenter and the plate. This method, capable of taking into account the high-velocity transverse (through-the-thickness) stress waves was subsequently extended to the calculation of damage zones in laminated composite plates subjected to low velocity impact (Bogdanovich & Yarve 1990). Sankar & Sun (1985b) used plane stress finite elements to study the low-velocity impact of laminated beams subjected to initial stresses.

Most of the impact problems have been formulated using the small deflection theory, which is adequate if the impact load is small. However, if a plate undergoes large deflections of the order of the thickness of the plate, it is necessary to include the geometric nonlinearity. Chen & Sun (1985) investigated the nonlinear transient response of composite laminates subjected to impact loads with initial stresses. They used the finite element method based on the Mindlin plate theory in conjunction with an experimentally established contact law (Tan & Sun 1982). Kant & Mallikarjuna (1991) used a higher-order theory and  $C^0$  finite elements to analyse a laminated plate under impact loads using the modified Hertzian contact law by Tan & Sun (1982). Obst & Kapania (1992) studied the geometrically nonlinear impact response of laminated beams using a third-order shear deformation theory. These authors also considered the effect of initial stresses. Effect of geometric imperfections on the geometrically nonlinear impact response of thin laminated plates was studied by Byun & Kapania (1992).

It is important to predict the combined effect of various damage modes and the external environment on the degradation and failure of the laminate. However, it appears that not much work has been done to understand and quantify the combined effect of these damage modes on the performance of composite laminates, whereas this understanding is essential for efficient design and production of damage resistant and durable laminated composite structures under impact loads. Hence, there is a need for developing a general tool to predict the combined effect of various damage modes on the performance of the composite structures with complex geometry and loading conditions.

To study the impact-generated damage, it is important to get very accurate information for in-plane and transverse stresses. However, finite element models based on the classical laminated plate theory (CLPT) or the first order shear deformation theory (FSDPT) cannot give accurate interlaminar stresses from constitutive relations. The equilibrium equations of 3-D elasticity give accurate interlaminar stresses. A post-processor for transverse normal and shear stresses as well as in-plane stresses was developed by Byun & Kapania (1991) using the finite element codes employing the CLPT and first order shear deformation theory. The transverse stresses were obtained by integrating the 3-D elasticity equilibrium equations. The postprocessor uses global interpolation of the nodal displacements obtained from the finite element analysis. Using a finite elements program based on both CLPT (Kapania & Yang 1986) and FSDPT (Reddy 1980), the transverse stresses were obtained for both symmetrically and unsymmetrically laminated plates. Good agreement with the 3-D elasticity results given by Pagano (1969) for symmetrically laminated plates and by Chaudhuri & Seide (1987) for unsymmetrically laminated plates was obtained. This formulation did not include the inertia effects.

An alternate way to accurately calculate the in-plane and transverse stresses,



without directly integrating the equilibrium equations, is to use Reddy's layerwise theory (Reddy 1987). One of the objectives of the present work is to use this theory for studying the impact response of laminated plates and to compare its performance to that of Reddy's first-order (Reddy 1983) and third-order (Reddy 1990) shear deformation theories. The layerwise theory of Reddy is expected to be more efficient than performing a conventional three-dimensional finite element analysis of impact response.

Recent experimental studies by Qian & Swanson (1990) and Swanson *et al* (1991) have shown that to accurately calculate the stresses in the vicinity of the impact, it is important to account for the distribution of the impact load. Note that it is very important that the stresses under the impactor be obtained very accurately as the matrix failure may initiate under the point of impact (Pintado *et al* 1991). In the present study, six different distributions, labelled as cases I through VI (with varying complexity), were conceived and the expressions for the generalized forces (to be used in the modal superposition technique) corresponding to each distribution were derived.

The first three models, in which the impact load is presented as (i) a concentrated load (model I), (ii) a uniform load distributed over a rectangular contact region (model II), and (iii) a sinusoidally distributed load (model III) over a rectangular contact region, are already considered in the literature. In model IV the load is assumed to be uniformly distributed over a circular region. This model is implemented in finite element codes by various researchers. Here, however, we considered it in our analytical solutions. In model V the impact load is assumed to be spherically distributed over a circular region in accordance with the Hertzian contact law. The analytical expression for the generalized forces corresponding to this load is also obtained for the first time in this paper. Despite the fact that the size of the contact region varies with time, it is assumed that the size of the contact region is known in cases I through V loading models. For models I through V, we obtain a nonlinear integral equation, whose solution yields time histories of the impact force, displacement, and other response quantities. Finally in model VI we incorporated certain relations, from the Hertzian contact law, into the case V loading model and obtained a slightly different nonlinear integral equation in which the time variance of the contact region is taken into account.

An iterative scheme with a small time increment is often used to obtain the response time-histories. Sankar & Sun (1985a) developed a numerical scheme in which the impact force is represented by a series of ramp functions. These ramp functions increase linearly from zero to unity during a time step and retain this value for the subsequent duration. Using a series of these ramp functions, one can calculate the value of the dynamic Green's function for a given structure. The advantage of this approach was that the iterations for solving the nonlinear equation for calculating the impact force were separated from the structural response calculations. In the present study, a different approach is proposed. The impact force is assumed to be represented by a series of piecewise basis functions (finite elements) in time. As a first step, the impact force is represented as a series of piecewise linear Lagrange interpolation functions (the so-called hat functions). An obvious advantage of this approach is its modularity, i.e., one can easily replace the linear interpolating polynomials by higher-order functions. Indeed, in this study we have also represented the impact force by using the time-finite-elements using cubic Hermite interpolating polynomials with minimal changes in the overall analysis.

## 2. Forced-vibration analysis: General results

Owing to the complex nature of dynamic load transfer characteristics, the response analysis of an impacted plate is more complicated than an ordinary dynamic problem in which the distribution of the transverse load is known. That is, the solution of a nonlinear integral equation is necessitated which, on the other hand, will result in the simultaneous knowledge of impact force and response quantities of a plate. For most part, however, the usual procedures used in an ordinary response analysis calculations can be utilised to determine the general analytical expressions for the response quantities in terms of the unknown impact force.

The systems of the differential eigenvalue equations of a laminated plate using the layerwise theory, presented in a report (Nosier *et al* 1992), can be shown to be self-adjoint. Therefore, we can make a positive statement about the orthogonality of the eigenfunctions, provided the eigenfrequencies (i.e. natural frequencies) are distinct. As a result, modal analysis can be conveniently used to obtain the response.

### 2.1 Orthogonality relationships of normal modes

The orthogonality relationships in various plate theories can be obtained by the use of standard procedures in the analysis of continuous systems. In the layerwise plate theory (LWPT), for example, if we let  $U_{mnk}^j$ ,  $V_{mnk}^j$ , and  $W_{mnk}^j$  (see Nosier *et al* 1992) and  $U_{pqr}^i$ ,  $V_{pqr}^i$ , and  $W_{pqr}^i$  denote the eigenfunctions corresponding to distinct eigenfrequencies  $\omega_{mnk}$  and  $\omega_{pqr}$ , respectively, the orthogonality relation can be stated as:

$$\int_0^b \int_0^a [I^{ij}(U_{mnk}^j U_{pqr}^i + V_{mnk}^j V_{pqr}^i + W_{mnk}^j W_{pqr}^i)] dx dy = \begin{cases} 0, & \text{if } m \neq p \text{ and/or } n \neq q \text{ and/or } k \neq r, \\ N_{mnk}, & \text{if } m = p, n = q, \text{ and } k = r. \end{cases} \quad (1)$$

Similarly, in third-order shear deformation theory (TSDPT,  $\lambda = 1$ ) and first-order shear deformation theory (FSDPT,  $\lambda = 0$ ) we have:

$$\begin{aligned} \int_0^b \int_0^a & \left[ I_1(U_{mnk} U_{pqr} + V_{mnk} V_{pqr} + W_{mnk} W_{pqr}) \right. \\ & + \bar{I}_2(U_{mnk} \Psi_{x_{pqr}} + \Psi_{x_{mnk}} U_{pqr} + V_{mnk} \Psi_{x_{pqr}} + \Psi_{y_{mnk}} V_{pqr}) \\ & + \bar{I}_3(\Psi_{x_{mnk}} \Psi_{x_{pqr}} + \Psi_{y_{mnk}} \Psi_{y_{pqr}}) \\ & - \lambda \frac{4}{3h^2} I_4(U_{mnk} W_{pqr,x} + W_{mnk,x} U_{pqr} + V_{mnk} W_{pqr,y} + W_{mnk,y} V_{pqr}) \\ & - \lambda \frac{4}{3h^2} \bar{I}_5(\Psi_{x_{mnk}} W_{pqr,x} + W_{mnk,x} \Psi_{x_{pqr}} + \Psi_{y_{mnk}} W_{pqr,x} + W_{mnk,y} \Psi_{y_{pqr}}) \\ & \left. + \lambda \left( \frac{4}{3h^2} \right)^2 I_7(W_{mnk,x} W_{pqr,x} + W_{mnk,y} W_{pqr,y}) \right] dx dy \\ & = \begin{cases} 0, & \text{if } m \neq p \text{ and/or } n \neq q \text{ and/or } k \neq r, \\ N_{mnk}, & \text{if } m = p, n = q, \text{ and } k = r, \end{cases} \end{aligned} \quad (2)$$

and in CLPT

$$\int_0^b \int_0^a I_1 (U_{mnk} U_{pqr} + V_{mnk} V_{pqr} + W_{mnk} W_{pqr}) dx dy = \begin{cases} 0, & \text{if } m \neq p \text{ and/or } n \neq q \text{ and/or } k \neq r, \\ N_{mnk}, & \text{if } m = p, n = q, \text{ and } k = r. \end{cases} \quad (3)$$

For a simply-supported plate the generalized masses  $N_{mnk}$  in various theories are given as (see Nosier et al 1992):

LWPT

$$N_{mnk} = (ab/4) I^{ij} (A_{mnk}^i A_{mnk}^j + B_{mnk}^i B_{mnk}^j + C_{mnk}^i C_{mnk}^j). \quad (4)$$

TSDPT ( $\lambda = 1$ ), FSDPT ( $\lambda = 0$ )

$$N_{mnk} = (ab/4) [I_1 (A_{mnk}^2 + B_{mnk}^2 + E_{mnk}^2) + 2\bar{I}_2 (A_{mnk} C_{mnk} + B_{mnk} D_{mnk}) + \bar{I}_3 (C_{mnk}^2 + D_{mnk}^2) - \lambda(8/3h^2) I_4 (\alpha_m A_{mnk} E_{mnk} + \beta_n B_{mnk} E_{mnk}) - \lambda(8/3h^2) \bar{I}_5 (\alpha_m C_{mnk} + \beta_n D_{mnk}) E_{mnk} + \lambda(4/3h^2)^2 I_7 (\alpha_m^2 + \beta_n^2) E_{mnk}^2]. \quad (5)$$

CLPT

$$N_{mnk} = (ab/4) I_1 (A_{mnk}^2 + B_{mnk}^2 + E_{mnk}^2). \quad (6)$$

Note that a repeated index in (4) indicates summation of terms over the range of that index.

## 2.2 Response analysis, general results

As we pointed out earlier, the response of the laminate to transverse excitation can be obtained conveniently by modal analysis. To this end, we illustrate the analysis within LWPT and summarize the appropriate results of the equivalent single-layer theories. According to normal mode analysis, we expand the primary response quantities as

$$u^j(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} U_{mnk}^j(x, y) \cdot \zeta_{mnk}(t), \quad (7a)$$

$$v^j(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} V_{mnk}^j(x, y) \cdot \zeta_{mnk}(t), \quad (7b)$$

$$w^j(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} W_{mnk}^j(x, y) \cdot \zeta_{mnk}(t), \quad j = 1, 2, \dots, N+1, \quad (7c)$$

where  $W_{mnk}^j$ ,  $V_{mnk}^j$ , and  $U_{mnk}^j$  are the eigenfunctions and  $\zeta_{mnk}$  are the time-dependent generalized coordinates, yet to be determined.

Upon substitution of (7) into the equations of motion we obtain (see Nosier et al 1992)

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} I^{ij} U_{mnk}^j (\ddot{\zeta}_{mnk} + \omega_{mnk}^2 \zeta_{mnk}) = 0, \quad (8a)$$

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} I^{ij} V_{mnk}^j (\zeta_{mnk}^i + \omega_{mnk}^2 \zeta_{mnk}^i) = 0, \quad (8b)$$

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} I^{ij} W_{mnk}^j (\zeta_{mnk}^i + \omega_{mnk}^2 \zeta_{mnk}^i) = \delta_{i1} P_z. \quad (8c)$$

Now we multiply (8a), (8b), and (8c), respectively, by  $U_{pqr}^i$ ,  $V_{pqr}^i$ ,  $W_{pqr}^i$  and add the results so obtained. We obtain

$$\begin{aligned} & \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} (\zeta_{mnk}^i + \omega_{mnk}^2 \zeta_{mnk}^i) \\ & \times [I^{ij} (U_{mnk}^j U_{pqr}^i + V_{mnk}^j V_{pqr}^i + W_{mnk}^j W_{pqr}^i)] = \delta_{i1} W_{pqr}^i P_z \equiv W_{pqr}^1 P_z. \end{aligned} \quad (9)$$

Integration of (9) over the surface of the plate and application of the orthogonality relation (1) results in

$$\zeta_{mnk}^i(t) + \omega_{mnk}^2 \zeta_{mnk}^i(t) = (1/N_{mnk}) \cdot Q_{mnk}(t), \quad (10a)$$

where the generalized forces  $Q_{mnk}$  are given by:

$$Q_{mnk}(t) = \int_0^b \int_0^a P_z \cdot W_{mnk}^1 dx dy. \quad (10b)$$

With the initial conditions being homogeneous, the formal solution of (10a) can be presented as

$$\zeta_{mnk}(t) = (1/\omega_{mnk} N_{mnk}) \int_0^t Q_{mnk}(t) \cdot \sin \omega_{mnk}(t - \tau) d\tau. \quad (11)$$

Similarly, the primary response quantities of the plate within the equivalent single-layer theories are obtained from

$$\begin{aligned} u(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K U_{mnk}(x, y) \cdot \zeta_{mnk}(t), \\ v(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K V_{mnk}(x, y) \cdot \zeta_{mnk}(t), \\ \psi_x(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^5 \Psi_{x_{mnk}}(x, y) \cdot \zeta_{mnk}(t), \\ \psi_y(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^5 \Psi_{y_{mnk}}(x, y) \cdot \zeta_{mnk}(t), \\ w(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K W_{mnk}(x, y) \cdot \zeta_{mnk}(t), \end{aligned} \quad (12)$$

where  $\zeta_{mnk}(t)$  are given by (11), with appropriate  $N_{mnk}$  in various theories, and  $Q_{mnk}(t)$  are defined as

$$Q_{mnk}(t) = \int_0^b \int_0^a P_z \cdot W_{mnk} dx dy. \quad (13)$$

In (12),  $K = 5$  in TSDPT and FSDPT and  $K = 3$  in CLPT.

Note that the assumption of transverse inextensibility is made in all the equivalent single-layer theories considered here. For this reason the local nature of the impact damage cannot be taken into account in such theories. On the other hand, comparing the expression of the generalized forces  $Q_{mnk}(t)$  in LWPT, (10b), with that in the equivalent single-layer theories (13), clearly reveals the appropriateness of LWPT for the impact problems.

We express  $N_{mnk}$  appearing in (4) as

$$N_{mnk} = C_{mnk}^1 \cdot C_{mnk}^1 \cdot \hat{N}_{mnk}, \quad (14)$$

where

$$\hat{N}_{mnk} = (ab/4) I^{ij} \left( \frac{A_{mnk}^i}{C_{mnk}^1} \frac{A_{mnk}^j}{C_{mnk}^1} + \frac{B_{mnk}^i}{C_{mnk}^1} \frac{B_{mnk}^j}{C_{mnk}^1} + \frac{C_{mnk}^i}{C_{mnk}^1} \frac{C_{mnk}^j}{C_{mnk}^1} \right).$$

Also the eigenfunctions can be presented as

$$\begin{aligned} U_{mnk}^j &= A_{mnk}^j \hat{U}_{mn}, \\ V_{mnk}^j &= B_{mnk}^j \hat{V}_{mn}, \\ W_{mnk}^j &= C_{mnk}^j \hat{W}_{mn}, j = 1, 2, \dots, N+1, \end{aligned} \quad (15)$$

where  $\hat{U}_{mn} = \cos \alpha_m x \sin \beta_n y$ ,  $\hat{V}_{mn} = \sin \alpha_m x \cos \beta_n y$ ,  $\hat{W}_{mn} = \sin \alpha_m x \sin \beta_n y$ ; and  $\alpha_m$  and  $\beta_n$  are defined in Nosier *et al* (1992).

With the help of (15), we can express the generalized forces  $Q_{mnk}(t)$  as

$$Q_{mnk}(t) = C_{mnk}^1 \hat{Q}_{mn}(t), \quad (16)$$

where  $\hat{Q}_{mn}(t) = \int_0^b \int_0^a P_z \hat{W}_{mn}(x, y) dx dy$ .

Introducing (14) and (16) into (11) results in

$$\zeta_{mnk}(t) = \frac{1}{\omega_{mnk}} \frac{1}{C_{mnk}^1} \frac{1}{\hat{N}_{mnk}} \cdot \hat{\zeta}_{mnk}(t), \quad (17)$$

where

$$\hat{\zeta}_{mnk}(t) = \int_0^t \hat{Q}_{mn}(\tau) \cdot \sin \omega_{mnk}(t - \tau) d\tau. \quad (18)$$

Finally, upon substitution of (15) and (17) into (7) we obtain

$$\begin{aligned} u^j(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} \frac{1}{\omega_{mnk} \hat{N}_{mnk}} \cdot \left( \frac{A_{mnk}^j}{C_{mnk}^1} \right) \cdot U_{mn}(x, y) \cdot \hat{\zeta}_{mnk}(t), \\ v^j(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} \frac{1}{\omega_{mnk} \hat{N}_{mnk}} \cdot \left( \frac{B_{mnk}^j}{C_{mnk}^1} \right) \cdot V_{mn}(x, y) \cdot \hat{\zeta}_{mnk}(t), \\ w^j(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} \frac{1}{\omega_{mnk} \hat{N}_{mnk}} \cdot \left( \frac{C_{mnk}^j}{C_{mnk}^1} \right) \cdot W_{mn}(x, y) \cdot \hat{\zeta}_{mnk}(t). \end{aligned} \quad (19)$$

Equations (19), in conjunction with (14), (15), (16), and (18), constitute the complete solution of the forced-vibration problem of a plate within LWPT.

Similarly, following the same line of reasoning, the primary response quantities of the plate according to various equivalent single-layer theories, as given by (12), can be obtained (see Nosier *et al* 1992).

### 2.3 Representation of impact load

Determination of impact induced surface pressure and its spatial distribution is a major task in impact analysis. Often, the functional dependency of the pressure on temporal and spatial coordinates is assumed to be separable as

$$P_z(x, y, t) = F(t) \cdot D(x, y), \quad (20)$$

where  $F(t)$  denotes the time-dependent amplitude of the pressure and the function  $D(x, y)$  describes the spatial distribution of the impact pressure in contact area. The amplitude  $F(t)$  is related to the indentation of the plate at the contact point by a contact law.

Here we describe several models for presenting the spatial distribution of the pressure  $P_z$  and evaluate the generalized forces  $\hat{Q}_{mn}(t)$  in each model. Later, it will be demonstrated that the time history of the plate deflection is virtually independent of  $D(x, y)$  as long as the total impulse, transmitted to the plate by the impactor, remains the same. The stress field in the contact region and in the vicinity of the contact area, is, however, very sensitive to the particular choice of pressure distribution.

*Case-I – loading model:* Assuming that the impact load can be described by a concentrated load at the point of contact, we have

$$P_z(x, y, t) = F(t) \cdot \delta(x - a_0, y - b_0), \quad (21)$$

where the generalized function  $\delta$  is the two-dimensional Dirac's delta function. The generalized force  $\hat{Q}_{mn}(t)$  is determined by substituting (15) and (21) into (16):

$$\hat{Q}_{mn}(t) = F(t) \cdot K_{mn} \cdot H_{mn}, \quad (22)$$

where  $K_{mn} = \sin \alpha_m \alpha_n \cdot \sin \beta_m \beta_n$ , and  $H_{mn} = 1$ .

Modelling the impact load as a concentrated load will result in an infinite shear force at the contact point in the plate. In cases II through V, it is assumed that the load is distributed over a known small rectangular or circular area. Further, its distribution will be uniform (cases II and IV), sinusoidal (case III), or spherical (case V).

*Case-II – loading model:* Assume that the impact load is uniformly distributed over a small rectangular area  $(\bar{a} \times \bar{b})$  as shown in figure 1. That is,

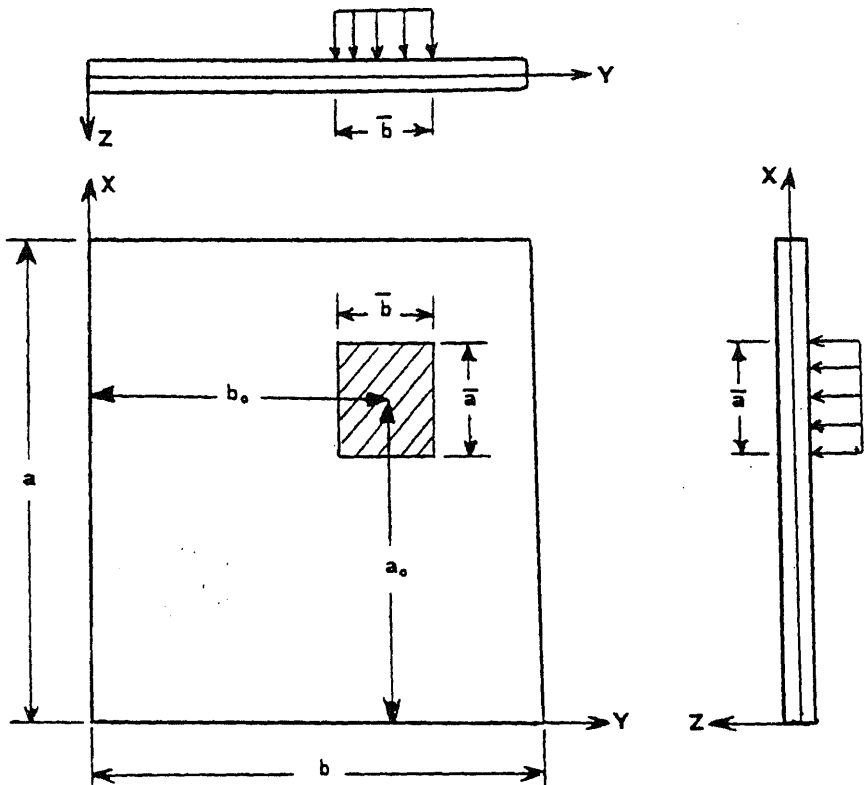
$$P_z(x, y, t) = \begin{cases} \frac{F(t)}{\bar{a}\bar{b}}, & a_0 - \frac{\bar{a}}{2} \leq x \leq a_0 + \frac{\bar{a}}{2}, b_0 - \frac{\bar{b}}{2} \leq y \leq b_0 + \frac{\bar{b}}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (23)$$

Evaluating  $\hat{Q}_{mn}(t)$  and expressing it as in (22), we have

$$H_{mn} = \left( \sin \alpha_m \frac{\bar{a}}{2} / \alpha_m \frac{\bar{a}}{2} \right) \left( \sin \beta_n \frac{\bar{b}}{2} / \beta_n \frac{\bar{b}}{2} \right). \quad (24)$$

In cases I through V,  $K_{mn}$  will be as given earlier.

*Case-III – loading model:* Here we assume that a cosine-shaped load is distributed



**Figure 1.** The impact load modelled as a uniform load distributed over a small rectangular region ( $\bar{a} \times \bar{b}$ ); case-II loading model.

over a small rectangular area. Hence

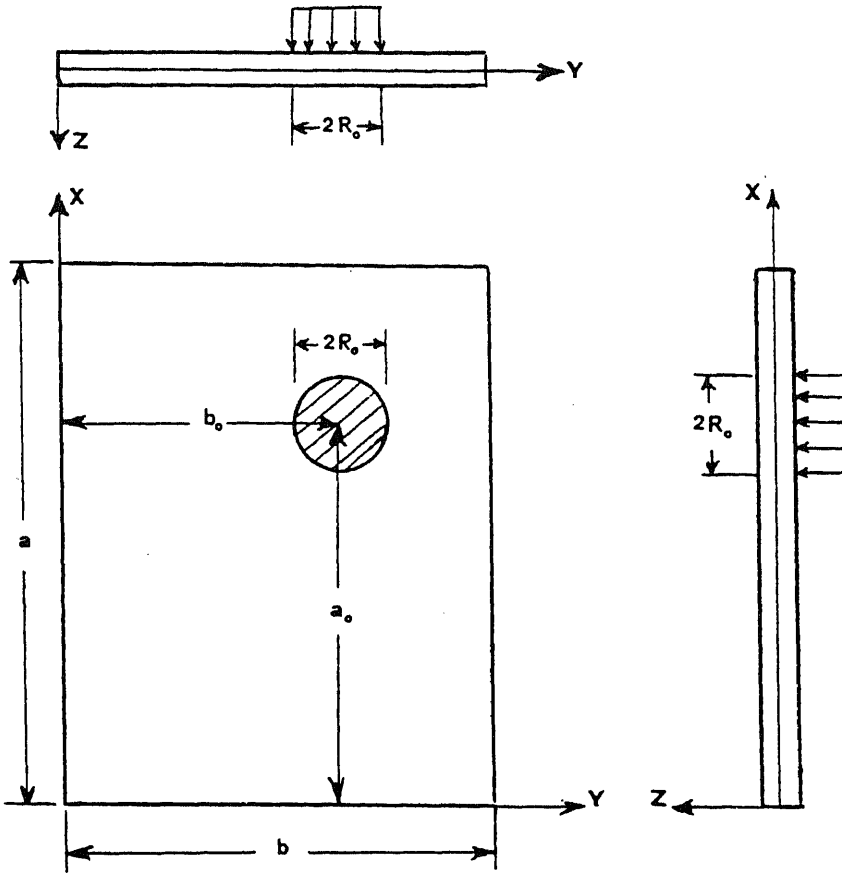
$$P_z(x, y, t) = \begin{cases} F(t) \cdot \frac{\pi^2}{4\bar{a}\bar{b}} \cdot \cos \frac{\pi}{\bar{a}}(x - a_0) \cdot \cos \frac{\pi}{\bar{b}}(y - \bar{y}), & a_0 - \frac{\bar{a}}{2} \leq x \leq a_0 + \frac{\bar{a}}{2} \\ & b_0 - \frac{\bar{b}}{2} \leq y \leq b_0 + \frac{\bar{b}}{2}, \\ 0, & \text{otherwise,} \end{cases} \quad (25)$$

and

$$H_{mn} = \left\{ \cos \alpha_m \frac{\bar{a}}{2} / \left[ 1 - \left( \alpha_m \frac{\bar{a}}{\pi} \right)^2 \right] \right\} \cdot \left\{ \cos \beta_n \frac{\bar{b}}{2} / \left[ 1 - \left( \beta_n \frac{\bar{b}}{\pi} \right)^2 \right] \right\}. \quad (26)$$

*Case-IV – loading model:* It is more realistic to model the contact area as a circular region, at least for an isotropic plate. Also, in the case of orthotropic plates the contact area is only slightly elliptical (see Greszczuk 1982). As a first approximation we assume that the impact load is uniformly distributed over this circular contact area, see figure 2. That is,

$$P_z(x, y, t) = \begin{cases} F(t)/\pi R_0^2, & \text{over the shaded area,} \\ 0, & \text{otherwise,} \end{cases} \quad (27)$$



**Figure 2.** The impact load modelled as a uniform load distributed over a small circular region; case-IV loading model.

and, therefore, (see appendix D in Nosier *et al* 1992)

$$H_{mn} = 2(J_1(\bar{R}))/\bar{R}, \quad (28)$$

where  $\bar{R} = (\alpha_m^2 + \beta_n^2)^{1/2} R_o$ ; and  $J_1$  is the Bessel function of the first kind and of order one.

**Case-V—loading model.** In the Hertzian contact law, the distribution of the contact load in the circular contact area is found to be spherically shaped (Timoshenko & Goodier 1970). It is to be remembered, however, that in the Hertz model both the target and the projectile were assumed to be isotropic and elastic; the target was assumed semi-infinite. For a spherically-shaped load we have (see figure 3):

$$P_z(x, y, t) = \begin{cases} F(t) \frac{3}{2\pi R_o^2} \left[ 1 - \frac{1}{R_o^2} (\bar{x}^2 + \bar{y}^2) \right]^{1/2}, & \text{in the shaded area,} \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$



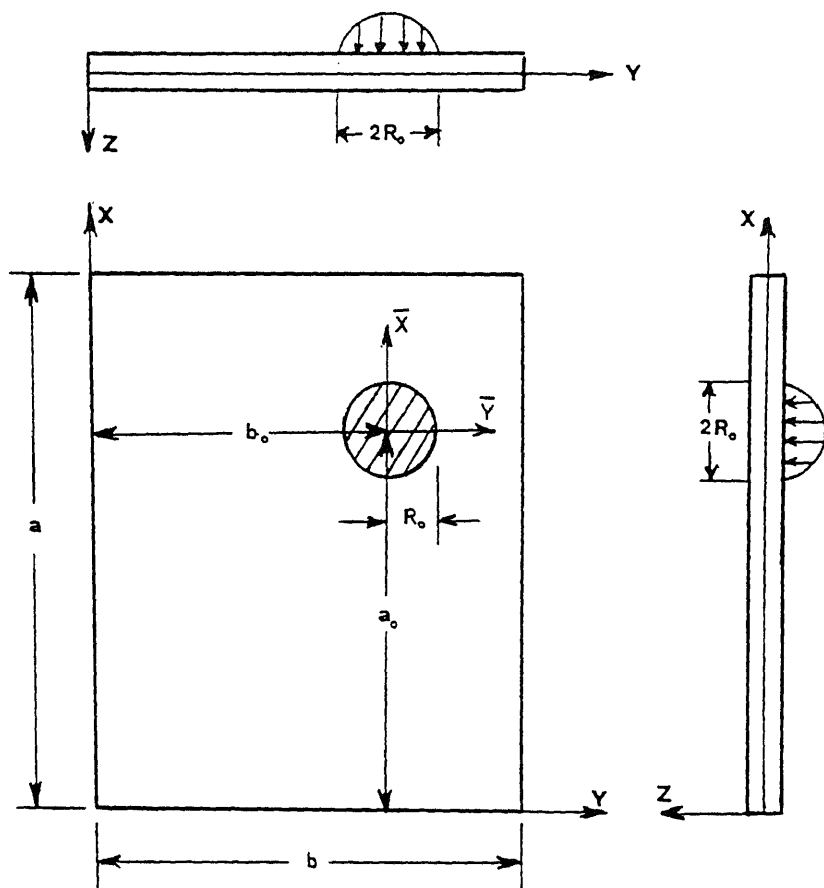


Figure 3. The impact load modelled as a spherically-shaped load over a small circular region; case-V loading model.

In this case we have (see appendix D in Nosier *et al* 1992):

$$H_{mn} = (3/\bar{R}^2)[(\sin \bar{R})/\bar{R} - \cos \bar{R}], \quad (30)$$

where  $\bar{R}$  is as defined earlier in conjunction with (28).

Having expressed the generalized forces in terms of the unknown amplitude of the impact load, the displacement of the plate at the contact point ( $x = a_o, y = b_o$ ) according to, for example, LWPT, is obtained by substituting (22) into (18) and the result so obtained into (19), with  $j = 1$ :

$$u_3(a_o, b_o, -h/2, t) = w^1(a_o, b_o, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} \bar{B}_{mnk} \int_0^t F(\tau) \sin \omega_{mnk}(t - \tau) d\tau, \quad (31)$$

where  $\bar{B}_{mnk} = (1/\omega_{mnk} N_{mnk}) K_{mn}^2 H_{mn}$ . Similarly, the plate displacement at the contact point according to HSDPT, FSDPT, and CLPT can be obtained (see Nosier *et al* 1992).

By invoking a contact law, the displacement and the amplitude of the impact force along with other pertinent response quantities can be determined.

*Case-VI – loading model – special treatment:* In cases II through V we have assumed that the impact pressure is distributed in a certain fashion over a known contact region. Generally, however, this contact area is unknown and changes with time. On the other hand, we mentioned that in the Hertz model, being based on completely elastic behaviour, the load distribution is spherically shaped. Furthermore, in Hertz's law of contact the magnitude of the total impact load  $F(t)$  is related to the indentation  $\alpha(t)$ , at the contact point, as

$$F(t) = K_2 \alpha^{3/2}(t), \quad (32)$$

with  $K_2$  being a constant coefficient.

We can relate the radius  $R_o$  of the circular contact area in the case-V loading to the force  $F(t)$  by using the relationship

$$R_o(t) = R_s^{1/2} \alpha^{1/2}(t), \quad (33)$$

which is valid for elastic deformation (Hertzian) law. That is, elimination of  $\alpha(t)$  from (33), with the help of (32), results in

$$R_o(t) = (R_s^{1/2}/K_2^{1/3}) \cdot F^{1/3}(t), \quad (34)$$

where  $R_s$  denotes the radius of the spherical impactor.

Upon substitution of (34) into (30) and the result so obtained along with (22) into (18), we obtain

$$\begin{aligned} \hat{\zeta}_{mnk}(t) = (3k_{mn}/\gamma_{mn}^3) \int_0^t \{ \sin[\gamma_{mn} F^{1/3}(\tau)] - \gamma_{mn} F^{1/3}(\tau) \cdot \cos[\gamma_{mn} F^{1/3}(\tau)] \} \\ \times \sin \omega_{mnk}(t - \tau) d\tau, \end{aligned} \quad (35)$$

where  $\gamma_{mn} = (\alpha_m^2 + \beta_n^2)^{1/2} (R_s^{1/2}/K_2^{1/3})$ . The displacement of the plate at the contact point is found by substituting (35) in (19), with  $j = 1$ :

$$\begin{aligned} u_3(a_o, b_o, -h/2, t) = w^1(a_o, b_o, t) \\ = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^{3(N+1)} \bar{B}_{mnk} \int_0^t \{ \sin[\gamma_{mn} F^{1/3}(\tau)] - \gamma_{mn} F^{1/3}(\tau) \cdot \cos[\gamma_{mn} F^{1/3}(\tau)] \} \\ \times \sin \omega_{mnk}(t - \tau) d\tau. \end{aligned} \quad (36)$$

Similarly, the displacement at the contact point according to the equivalent single-layer theories is given by:

$$\begin{aligned} u_3(a_o, b_o, -h/2, t) = w(a_o, b_o, t) \\ = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \int_0^t \{ \sin[\gamma_{mn} F^{1/3}(\tau)] - \gamma_{mn} F^{1/3}(\tau) \cdot \cos[\gamma_{mn} F^{1/3}(\tau)] \} \\ \times \sin \omega_{mnk}(t - \tau) d\tau \end{aligned}$$

where  $\bar{B}_{mnk} = 3K_{mn}^2/(\gamma_{mn}^3 \omega_{mnk} \hat{N}_{mnk})$  in all theories, where the  $\hat{a}$  for  $\hat{N}_{mnk}$  must be used.

In the case-VI model, the contact area, being time loading process from zero to a maximum when the force

then decreases in the unloading process until it becomes zero again at the time of rebound.

In the next section, two algorithms will be introduced for the determination of the impact force and responses quantities for loading cases I through V. The determination of impact force and response of the plate for the case-VI loading model will be followed in a separate section. The variation of the total impact force  $F(t)$  and the transverse displacement of the laminate will, of course, be approximately identical for all the loading models we are considering here. This conclusion is also valid as far as the stress field away from the contact point is concerned. The significance of the various models will become apparent only when we consider the stress field in the vicinity of the impact zone.

### 3. Impact response. Analysis

The direct measurement of the dynamic contact force is a complicated task due to the existence of a wide range of plate and impactor parameters. It will, therefore, be assumed that the vibrations of the impactor and the laminate can be neglected so that an elastostatic indentation law can be used. That is, the impact duration will be assumed to be long compared to the stress-wave transit times in the impactor and the laminate.

#### 3.1 Contact laws

The most famous elastic contact law was derived by Hertz (see (32)) for the contact of two spheres of elastic isotropic materials. The contact between a sphere on a half-space is a limiting case. In (32)  $\alpha$  denotes the indentation and the contact coefficient  $K_2$  is given by

$$K_2 = (4/3)[R_s^{1/2}/(\delta_1 + \delta_2)], \quad (38)$$

where  $\delta_1 = (1 - \nu_s^2)/E_s$ ; and  $\delta_2 = (1 - \nu^2)/E$ . In (38),  $R_s$  denotes the radius of the spherical impactor. Also  $(\nu_s, E_s)$  and  $(\nu, E)$  are the Poisson ratio and Young's modulus of the impactor and the half-space, respectively. The indentation law for an isotropic spherical impactor and a transversely isotropic half-space (see Conwag 1956) is also given by (32). The  $x$ - $y$  plane is assumed to be the plane of isotropy. The expressions of  $K_2$  and  $\delta_1$  are again given as mentioned earlier in conjunction with (38), and

$$\delta_2 = \bar{A}_{22}^{1/2} \{ [(\bar{A}_{11} \bar{A}_{22})^{1/2} + G_z]^2 - (\bar{A}_{12} + G_z)^2 \}^{1/2} \\ \div \{ 2G_z^{1/2} (\bar{A}_{11} \bar{A}_{22} - \bar{A}_{22}^2) \} \quad (39)$$

where

$$\bar{A}_{11} = E_z / \{ 1 - [2\nu_z^2 E_z / (1 - \nu) E] \}, \quad \bar{A}_{22} = [(E/E_z) - \nu_z^2] \bar{A}_{11} / (1 - \nu_z^2), \\ \bar{A}_{12} = \nu_z \bar{A}_{11} / (1 - \nu).$$

Here,  $\nu$  and  $E$  are the Poisson ratio and Young's modulus in the plane of isotropy ( $x$ - $y$  plane), respectively. That is,  $\nu = \nu_{12} = \nu_{21}$  and  $E = E_1 = E_2$ . Also,  $\nu_z = \nu_{13} = \nu_{23}$  and  $E_z = E_3$  with the shear modulus  $G_z = G_{13} = G_{23}$ .

When the plate is laminated of orthotropic layers, Sun (1977) employed the Hertz law with  $\delta_2 = 1/E_2$  where  $E_2$  is the Young modulus transverse to the fibre direction

in the  $x-y$  plane. The value of  $K_2$  so determined was observed by Yang & Sun (1981) to be higher than that obtained experimentally for a cross-ply laminate. At the present time the exact value of  $\delta_2$  for an orthotropic half-space is not known. However, the approximate numerical solution for  $\delta_2$  shows that  $\delta_2$  is relatively insensitive to the in-plane fibre orientation (Greszczuk 1975). For this reason, it is concluded by Greszczuk (1982) that for an orthotropic target, (39) can be used, as a first approximation, if average in-plane properties, say,  $\nu = (\nu_{12} + \nu_{21})/2$ ,  $E = (E_1 + E_2)/2$  are used. It is, however, noted by Greszczuk (1975) that the properties that influence  $\delta_2$  the most are those associated with the thickness ( $z$  -) direction. When an isotropic sphere is pressed into either an isotropic or transversely isotropic half-space, the area of contact is circular according to elasticity solutions. Furthermore, the radius  $R_0$  of this area is related to the radius  $R_s$  of the sphere and the indentation  $\alpha$  according to (see Timoshenko & Goodier 1970):  $R_0 = R_s^{1/2} \alpha^{1/2}$ .

The Hertzian contact law, being an elastic law, does not account for permanent indentation. Permanent indentation may often take place even at relatively low loading levels in composite targets and, therefore, the unloading curve may in general be different from the loading curve. The experimental studies of Tan & Sun (1982) indicated that the following approximate relations can be used as a proper contact law for composite targets:

$$\text{loading:} \quad F = K_2 \alpha^{3/2}, \quad (40a)$$

$$\text{unloading:} \quad F = F_m [(\alpha - \alpha_o)/(\alpha_m - \alpha_o)]^q, \quad (40b)$$

$$\text{reloading:} \quad F = F_m [(\alpha - \alpha_o)(\alpha_m - \alpha_o)]^{3/2}, \quad (40c)$$

where

$$\alpha_o = \begin{cases} \beta(\alpha_m - \alpha_p), & \text{if } \alpha_m > \alpha_p, \\ 0, & \text{if } \alpha_m < \alpha_p. \end{cases} \quad (40d)$$

In (40),  $F_m$  denotes the maximum force reached before unloading,  $\alpha_m$  is the corresponding indentation, and  $\alpha_o$  is the permanent indentation. The values of  $K_2$ ,  $\beta$ ,  $q$ , and  $\alpha_p$  are found experimentally. As a first approximation, however,  $K_2$  can be determined from (38), with  $\delta_2 = 1/E_2$ .

### 3.2 Determination of impact force

In this section we are concerned with the determination of the impact force when the transverse load is represented as any one of the cases I through V loading models. In determining the impact force we follow the basic approach developed by Timoshenko (1913), who studied the impact of an isotropic beam by a sphere (also see Goldsmith 1960). Denoting by  $m_0$ ,  $v_0$ , and  $z$  the mass of the impactor, the velocity of the impactor at the moment of impact, and the displacement of the impactor after the impact, we have

$$z(t) = v_0 t - (1/m_0) \int_0^t F(\tau)(t - \tau) d\tau. \quad (41)$$

Furthermore, if impact on the laminate occurs at  $(x = a_0, y = b_0)$ , we have figure 4):

$$\alpha(t) = z(t) - u_3(a_0, b_0, -h/2, t),$$

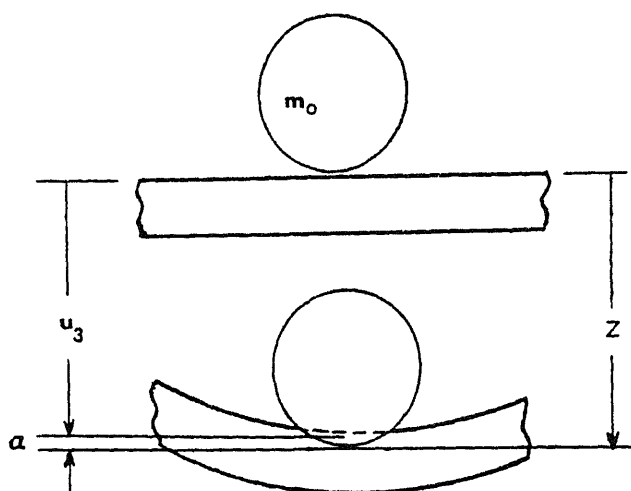


Figure 4. A plate before and after impact.

where  $u_3(a_0, b_0, -h/2, t)$  is the displacement of the laminate, at the point of impact, whose expression is given earlier. Using Hertz contact law (see Nosier *et al* 1992), we obtain

$$K_2^{-2/3} F^{2/3}(t) = v_0 t - (1/m_0) \int_0^t F(\tau)(t - \tau) d\tau - \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \int_0^t F(\tau) \sin \omega_{mnk}(t - \tau) d\tau, \quad (43)$$

where  $\bar{B}_{mnk}$  and  $K$  have been defined earlier. Upon solving (43) the time history of the impact force will be obtained. Here we use two different numerical schemes to determine the approximate values of the impact force at various time steps. Our first approach will be to assume a linear variation for the force during each small time increment. This approach will be very similar to the one presented by Sankar & Sun (1985a) but it will have one essential difference. Sankar & Sun (1985a) have taken the force variation as

$$F(t) = \sum_{i=0,1,2,\dots} q_i R\langle t - i\Delta t \rangle, \quad (44)$$

where the  $q_i$ 's are unknowns that decide the contact force history, and  $R\langle t - i\Delta t \rangle$  is the function defined as follows:

$$R\langle t - t_0 \rangle = \begin{cases} 0, & \text{for } t \leq t_0, \\ (t - t_0)/\Delta t, & \text{for } t_0 \leq t \leq t_0 + \Delta t, \\ 1, & \text{for } t \geq t_0 + \Delta t. \end{cases} \quad (45)$$

In our first approach we alternatively assume that

$$F(t) = \sum_{i=1,2,\dots} f_i \Phi_i(t), \quad (46)$$

where  $\Phi_i$ 's are the linear Lagrange form interpolation polynomials (the so-called hat

functions). That is

$$\Phi_i(t) = \begin{cases} 0, & \text{for } t \leq (i-1)\Delta t, \\ \psi_i^1 = (1/\Delta t)[t(i-1)\Delta t], & \text{for } (i-1)\Delta t \leq t \leq i\Delta t, \\ \psi_i^2 = (1/\Delta t)[(i-1)\Delta t - t], & \text{for } i\Delta t \leq t \leq (i+1)\Delta t, \\ 0, & \text{for } t \geq (i+1)\Delta t. \end{cases} \quad (47)$$

From (46) and (47) at time  $t = i\Delta t$  we have  $F(i\Delta t) = f_i$ . Hence,  $f_1, f_2, \dots$  are the magnitudes of the total impact force at times equal to  $\Delta t, 2\Delta t, \dots$ , respectively. Also when  $(i-1)\Delta t \leq t \leq i\Delta t$ , we have

$$F(t) = [if_{i-1} - (i-1)f_i] + (1/\Delta t)(f_i - f_{i-1})t, \quad i = 1, 2, \dots, \quad (48)$$

with  $f_0 = F(0) = 0$ . Next we assume that  $t = p\Delta t$  ( $p = 1, 2, \dots$ ) and evaluate the two integrals appearing in (43), with the help of (48), and obtain

$$\int_0^t F(\tau)(t-\tau)d\tau = [(\Delta t)^2/6] \cdot f_p + \delta_p \sum_{i=1}^{p-1} (p-i) \cdot (\Delta t)^2 \cdot f_i \quad (49a)$$

and

$$\begin{aligned} \int_0^t F(\tau) \cdot \sin \omega_{mnk}(t-\tau)d\tau = & \left( \frac{1}{\omega_{mnk}} - \frac{1}{\Delta t \omega_{mnk}^2} \sin \omega_{mnk} \Delta t \right) \cdot f_p \\ & + \delta_p \sum_{i=1}^{p-1} \left[ \frac{2}{\Delta t \omega_{mnk}^2} (1 - \cos \omega_{mnk} \Delta t) \cdot \sin \omega_{mnk}(p-i)\Delta t \right] \cdot f_i \end{aligned} \quad (49b)$$

where  $\delta_p = 0$  for  $p = 1$  and  $\delta_p = 1$  for all other values of  $p$ . Also at time  $t = p\Delta t$  let  $F(p\Delta t) = f_p$ . Then from (43) and (49), we have:

$$a_p f_p^{2/3} + b_p f_p + c_p = 0, \quad (50)$$

where the constant coefficients  $a_p$ ,  $b_p$ , and  $c_p$  are given in Nosier *et al* (1992). Now by introducing the new variable  $g_p$  as  $f_p = g_p^3$  in (50) we obtain

$$b_p g_p^3 + a_p g_p^2 + c_p = 0, \quad (51)$$

whose roots can be found exactly, and the magnitudes of the total impact force  $F(t)$  at various times  $p\Delta t$  ( $p = 1, 2, \dots$ ) can be obtained.

When the experimental contact law of Tan & Sun (1982) (40) is employed, (50) will be valid for the loading process. However, using the expression for  $\alpha(t)$  from (40b) in (43) and carrying out the above analysis we found that for the unloading process the magnitudes of the total impact force at various times  $t = p\Delta t$  are found from

$$\bar{K}_3 f_p^{1/4} + b_p f_p + c_p + \alpha_0 = 0, \quad (52a)$$

where  $\bar{K}_3$ ,  $b_p$  and  $c_p$  are given by Nosier *et al* (1992). Similarly, the magnitudes of the impact force at various times are obtained from

$$\bar{K}_4 f_p^{2/3} + b_p f_p + c_p \alpha_0 = 0, \quad (52b)$$

where  $\bar{K}_4$  is given by Nosier *et al* (1992). It is to be noted that when  $\alpha_0 = 0$ , (52b) becomes identical to (50). That is, the reloading contact law of Tan & Sun (1982)

will be identical to the Hertzian contact law. The manner in which (52a) will be solved depends on the experimentally evaluated value of  $q$ . The experimental results of Tan & Sun (1982) indicated that this value varied between 2 and 3. For  $q = 2.5$ , (52a) will be solved numerically. However for  $q = 2$  and  $q = 3$  (52a) can be transformed, respectively, into a quadratic equation and a cubic equation which can be solved exactly.

The impactor position at any time  $t = p\Delta t$  can be obtained as:

$$z(p\Delta t) = v_0 \cdot (p\Delta t) - \frac{1}{6} \frac{(\Delta t)^2}{m_0} \cdot f_p - \frac{\delta_p}{m_0} \sum_{i=1}^{p-1} (p-i) \cdot (\Delta t)^2 \cdot f_i. \quad (53a)$$

The transverse displacement of the laminate at the point of contact according to the layerwise theory and the equivalent single-layer theories can be written as

$$\begin{aligned} u_3(a_0, b_0, -h/2, p\Delta t) = & \left[ \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \cdot \left( \frac{1}{\omega_{mnk}} - \frac{1}{\Delta t \omega_{mnk}^2} \sin \omega_{mnk} \Delta t \right) \right] \cdot f_p \\ & + \delta_p \sum_{i=1}^{p-1} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \left[ \frac{2}{\Delta t \omega_{mnk}^2} (1 - \cos \omega_{mnk} \Delta t) \cdot \sin \omega_{mnk} (p-i) \Delta t \right] \cdot f_i, \end{aligned} \quad (53b)$$

where the values of  $K$  for various theories have been given earlier.

Next, in the second scheme, we assume that the contact force variation is cubic in time during each time increment and employ the global Hermite cubic interpolation polynomials.

$$F(t) = \sum_{i=1,2,\dots} [f_i \Phi_i(t) + \bar{f}_i \bar{\Phi}_i(t)], \quad (54)$$

where  $\Phi_i$ 's and  $\bar{\Phi}_i$ 's are given as:

$$\Phi_i(t) = \begin{cases} 0, & \text{for } t \leq (i-1)\Delta t, \\ \psi_i^1(t), & \text{for } (i-1)\Delta t \leq t \leq i\Delta t, \\ \psi_i^2(t), & \text{for } i\Delta t \leq t \leq (i+1)\Delta t, \\ 0, & \text{for } t \geq (i+1)\Delta t, \end{cases} \quad (55a)$$

and

$$\bar{\Phi}_i(t) = \begin{cases} 0, & \text{for } t \leq (i-1)\Delta t, \\ \bar{\psi}_i^1(t), & \text{for } (i-1)\Delta t \leq t \leq i\Delta t, \\ \bar{\psi}_i^2(t), & \text{for } (i-1)\Delta t \leq t \leq (i+1)\Delta t, \\ 0, & \text{for } t \geq (i+1)\Delta t, \end{cases} \quad (55b)$$

where

$$\begin{aligned} \psi_i^1(t) &= (3/\Delta t^2)[t - (i-1)\Delta t]^2 - (2/\Delta t^3)[t - (i-1)\Delta t]^3, \\ \psi_i^2(t) &= 1 - (3/\Delta t^2)(t - i\Delta t)^2 + (2/\Delta t^3)(t - i\Delta t)^3, \\ \bar{\psi}_i^1(t) &= (1/\Delta t^2)[t - (i-1)\Delta t]^3 - (1/\Delta t)[t - (i-1)\Delta t]^2, \\ \bar{\psi}_i^2(t) &= (t - i\Delta t) - (2/\Delta t)(t - i\Delta t)^2 + (1/\Delta t^2)(t - i\Delta t)^3. \end{aligned} \quad (56)$$

In (56) we have

$$f_i = F(i\Delta t) \text{ and } \bar{f}_i = \left. \frac{dF}{dt} \right|_{t=i\Delta t} \quad (57)$$

That is,  $f_i$  is, as before, the magnitude of the total impact force at time  $t = i\Delta t$  and  $\bar{f}_i$  is the slope of the force-time curve at time  $t = i\Delta t$ . Furthermore, we have  $\dot{f}_0 = F(0) = 0$  and  $\bar{f}_0 = \left. \frac{dF}{dt} \right|_{t=0} = 0$ . The last result can be obtained by differentiation of (32) and noting that  $\alpha(0) = 0$ . So, in summary, we are required to determine  $f_i$  and  $\bar{f}_i$  at each time step ( $t = i\Delta t$ ). For this purpose we need another equation besides (43). This second equation is obtained by direct differentiation of (43):

$$(2/3)K_2^{-2/3}\dot{F}(t) \cdot F^{-1/3}(t) = v_0 - (1/m_0) \int_0^t F(\tau) d\tau - \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \omega_{mnk} \bar{B}_{mnk} \int_0^t F(\tau) \cdot \cos \omega_{mnk}(t - \tau) d\tau. \quad (58)$$

When  $(i-1)\Delta t \leq t \leq i\Delta t$ , we have

$$F(t) = f_{i-1} \psi_{i-1}^2 + \bar{f}_{i-1} \bar{\psi}_{i-1}^2 + f_i \psi_i^1 + \bar{f}_i \bar{\psi}_i^1, \quad i = 1, 2, \dots \quad (59)$$

We evaluate the four integrals appearing in (43) and (58) at time  $t = p\Delta t$  ( $p = 1, 2, \dots$ ):

$$\begin{aligned} \int_0^t F(\tau) \cdot (t - \tau) d\tau &= \int_0^{p\Delta t} F(\tau) \cdot (p\Delta t - \tau) d\tau \\ &= \frac{3}{20}(\Delta t)^2 \cdot f_p - \frac{1}{30}(\Delta t)^3 \cdot \bar{f}_p + \delta_p \sum_{i=1}^{p-1} (\Delta t)^2 \left[ (p-i) \cdot f_i - \frac{\Delta t}{15} \cdot \bar{f}_i \right], \end{aligned} \quad (60a)$$

$$\begin{aligned} \int_0^t F(\tau) \cdot \sin \omega_{mnk}(t - \tau) d\tau &= \int_0^{p\Delta t} F(\tau) \cdot \sin \omega_{mnk}(p\Delta t - \tau) d\tau \\ &= \left[ \frac{1}{\omega_{mnk}} - \frac{12}{\Delta t^3 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t + \frac{6}{\Delta t^2 \omega_{mnk}^3} (1 + \cos \omega_{mnk} \Delta t) \right] \cdot f_p \\ &\quad + \left[ \frac{6}{\Delta t^2 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t - \frac{2}{\Delta t \omega_{mnk}^3} (2 + \cos \omega_{mnk} \Delta t) \right] \cdot \bar{f}_p \\ &\quad + \delta_p \cdot \sum_{i=1}^{p-1} \left\{ \sin \omega_{mnk} \Delta t (p-i) \left[ \frac{24}{\Delta t^3 \omega_{mnk}^4} (1 - \cos \omega_{mnk} \Delta t) - \frac{12}{\Delta t^2 \omega_{mnk}^3} \sin \omega_{mnk} \Delta t \right] \right\} \cdot f_i \\ &\quad + \delta_p \cdot \sum_{i=1}^{p-1} \left\{ \cos \omega_{mnk} \Delta t (p-i) \left[ \frac{12}{\Delta t^2 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t - \frac{4}{\Delta t \omega_{mnk}^3} (2 + \cos \omega_{mnk} \Delta t) \right] \right\} \cdot \bar{f}_i, \end{aligned} \quad (60b)$$

$$\int_0^t F(\tau) d\tau = \int_0^{p\Delta t} F(\tau) d\tau = \frac{1}{2} \Delta t \cdot f_p - \frac{1}{12} (\Delta t)^2 \cdot \bar{f}_p + \delta_p \cdot \sum_{i=1}^{p-1} \Delta t \cdot f_i, \quad (60c)$$

$$\begin{aligned} \int_0^t F(\tau) \cdot \cos \omega_{mnk}(t - \tau) d\tau &= \int_0^{p\Delta t} F(\tau) \cdot \cos \omega_{mnk}(p\Delta t - \tau) d\tau \\ &= \left[ \frac{12}{\Delta t^3 \omega_{mnk}^4} (1 - \cos \omega_{mnk} \Delta t) - \frac{6}{\Delta t^2 \omega_{mnk}^3} \sin \omega_{mnk} \Delta t \right] \cdot f_p \end{aligned}$$



$$\begin{aligned}
& + \left[ \frac{1}{\omega_{mnk}^2} + \frac{2}{\Delta t \omega_{mnk}^3} \sin \omega_{mnk} \Delta t - \frac{6}{\Delta t^2 \omega_{mnk}^4} (1 - \cos \omega_{mnk} \Delta t) \right] \cdot \bar{f}_p \\
& + \delta_p \cdot \sum_{i=1}^{p-1} \left\{ \cos \omega_{mnk} \Delta t (p-i) \left[ \frac{24}{\Delta t^3 \omega_{mnk}^4} (1 - \cos \omega_{mnk} \Delta t) - \frac{12}{\Delta t^2 \omega_{mnk}^3} \sin \omega_{mnk} \Delta t \right] \right\} \cdot f_i \\
& + \delta_p \cdot \sum_{i=1}^{p-1} \left\{ \sin \omega_{mnk} \Delta t (p-i) \left[ \frac{4}{\Delta t \omega_{mnk}^3} (2 + \cos \omega_{mnk} \Delta t) - \frac{12}{\Delta t^2 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t \right] \right\} \cdot \bar{f}_i,
\end{aligned} \tag{60d}$$

where  $\delta_p$  has been defined earlier. From (57), (60), (43) and (58) we have:

$$a_p f_p^{2/3} + b_p f_p + c_p \bar{f}_p + d_p = 0, \tag{61a}$$

and

$$\bar{a}_p \bar{f}_p \cdot f_p^{-1/3} + \bar{b}_p f_p + \bar{c}_p \bar{f}_p + \bar{d}_p = 0, \tag{61b}$$

where the coefficients  $a_p, b_p, \dots, \bar{d}_p$  are given in appendix E in Nosier et al (1992). Introducing  $f_p = g_p^3$ , (61b) becomes:

$$h_{4p} \cdot g_p^4 + h_{3p} g_p^3 + h_{2p} g_p^2 + h_{1p} g_p + h_{0p} = 0. \tag{62}$$

The expressions for  $h_j (j=0, 1, \dots, 4)$  in terms of  $a_p, b_p, \dots, \bar{d}_p$  are displayed in appendix E in Nosier et al (1992).

Upon solving (62) for  $g_p$  (for example by the Newton-Raphson method),  $f_p$  and  $\bar{f}_p$  can be determined for every  $p = 1, 2, \dots$ . The impactor position at any time  $t = p\Delta t$  is obtained by:

$$z(p\Delta t) = v_o(p\Delta t) - \frac{3(\Delta t)^2}{20 m_0} f_p + \frac{1(\Delta t)^3}{30 m_0} \bar{f}_p - \delta_p \cdot \sum_{i=1}^{p-1} \frac{(\Delta t)^2}{m_0} \left[ (p-i) f_i - \frac{\Delta t}{15} \bar{f}_i \right]. \tag{63}$$

The displacement of the plate at the contact point ( $x = a_0, y = b_0$ ) according to various plate theories is obtained by substituting (60b) into (31):

$$u_3(a_0, b_0, -h/2, p\Delta t) =$$

$$\begin{aligned}
& \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \left\{ \left[ \frac{1}{\omega_{mnk}} - \frac{12}{\Delta t^3 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t + \frac{6}{\Delta t^2 \omega_{mnk}^3} (1 + \cos \omega_{mnk} \Delta t) \right] f_p \right. \\
& + \left[ \frac{6}{\Delta t^2 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t - \frac{2}{\Delta t \omega_{mnk}^3} (2 + \cos \omega_{mnk} \Delta t) \right] \bar{f}_p \\
& + \delta_p \sum_{i=1}^{p-1} \left\{ \sin \omega_{mnk} \Delta t (p-i) \left[ \frac{24}{\Delta t^3 \omega_{mnk}^4} (1 - \cos \omega_{mnk} \Delta t) - \frac{12}{\Delta t^2 \omega_{mnk}^3} \sin \omega_{mnk} \Delta t \right] \right\} f_i \\
& + \delta_p \sum_{i=1}^{p-1} \left\{ \cos \omega_{mnk} \Delta t (p-i) \left[ \frac{12}{\Delta t^2 \omega_{mnk}^4} \sin \omega_{mnk} \Delta t - \frac{4}{\Delta t \omega_{mnk}^3} (2 + \cos \omega_{mnk} \Delta t) \right] \right\} \bar{f}_i \Big\}.
\end{aligned} \tag{64}$$

where, as before,  $K = 3(N+1)$  in LWPT,  $K = 5$  in FSDPT and TSDPT, and  $K = 3$  in CLPT.

### 3.3 Determination of impact force using loading VI

Here we consider the determination of the total impact force according to the loading model VI. The transverse displacement of the laminate at the impact point according to various theories is determined from (35) and (37). Substituting these equations and (32) and (41) into (43) results in

$$K_2^{-2/3} F^{2/3}(t) = v_0 t - (1/m_0) \int_0^t F(\tau)(t - \tau) d\tau - \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \\ \times \int_0^t \{ \sin[\gamma_{mn} F^{1/3}(\tau)] - \gamma_{mn} F^{1/3}(\tau) \cos[\gamma_{mn} F^{1/3}(\tau)] \} \sin \omega_{mnk}(t - \tau) d\tau, \quad (65)$$

where  $\gamma_{mn}$  has been defined earlier. The total force  $F(t)$  in (65) can be replaced by either  $\alpha(t)$  with the help of (32) or by the radius of contact region  $R_o(t)$  with the help of (34). In either case, the nonlinear integral equation (65) may not have an analytical solution. Therefore, only a numerical solution of this equation will be presented here. Assuming a linear variation for the total force  $F(t)$  during each small time increment, the first integral in (65) at time  $t = p\Delta t$  ( $p = 1, 2, \dots$ ) is given as in (49a). To evaluate the second integral appearing in (65) we assume that the total force  $F(t)$  is constant during each small time increment. That is, for  $(i-1)\Delta t \leq t \leq i\Delta t$ , we assume that

$$F(t) = \frac{1}{2}(f_i + f_{i-1}) = \tilde{f}_i, \quad (66)$$

where, as before,  $f_i$  is the magnitude of the force  $F(t)$  at time  $t = i\Delta t$ . The variable  $\tilde{f}_i$  is introduced for simplicity. The second integral in (65) will be

$$\int_0^{p\Delta t} \{ \sin[\gamma_{mn} F^{1/3}(\tau)] - \gamma_{mn} F^{1/3}(\tau) \cos[\gamma_{mn} F^{1/3}(\tau)] \} \cdot \sin \omega_{mnk}(p\Delta t - \tau) d\tau \\ = \frac{1}{\omega_{mnk}} (1 - \cos \omega_{mnk} \Delta t) [\sin(\gamma_{mn} \tilde{f}_p^{1/3}) - \gamma_{mn} \tilde{f}_p^{1/3} \cdot \cos(\gamma_{mn} \tilde{f}_p^{1/3})] \\ + \frac{\delta_p}{\omega_{mnk}} \sum_{i=1}^{p-1} [\cos \omega_{mnk}(p-i)\Delta t - \cos \omega_{mnk}(p-i+1)\Delta t] \\ \times [\sin(\gamma_{mn} \tilde{f}_i^{1/3}) - \gamma_{mn} \tilde{f}_i^{1/3} \cos(\gamma_{mn} \tilde{f}_i^{1/3})]. \quad (67)$$

Upon substitution of (49a), and (67) into (65) we obtain

$$a_p(2\tilde{f}_p - f_{p-1})^{2/3} + b_p \tilde{f}_p + c_p \\ + \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \frac{\bar{B}_{mnk}}{\omega_{mnk}} (1 - \cos \omega_{mnk} \Delta t) \\ \times [\sin(\gamma_{mn} \tilde{f}_i^{1/3}) - \gamma_{mn} \tilde{f}_p^{1/3} \cdot \cos(\gamma_{mn} \tilde{f}_p^{1/3})] = 0, \quad (68)$$

where  $a_p$ ,  $b_p$ , and  $c_p$  are given in Nosier *et al* (1992).

Upon numerical solution of (68) the value of  $\tilde{f}_p$  at  $t = p\Delta t$  and, therefore, the value of  $f_p$ , will be obtained. The impactor position is determined from (63). The transverse displacement of the laminate at the point of impact according to various theories at time  $t = p\Delta t$  is obtained by substituting (67) into (36) and (37).

#### 4. Numerical results and discussions

Throughout our numerical examples we assume that the impact occurs at the middle of the plate (or beam). That is  $a_0 = a/2$  and  $b_0 = b/2$ .

##### Example 1

Timoshenko (1913) considered the problem of a simply-supported isotropic beam ( $1 \times 1 \times 30.7$  cm) subjected to the transverse impact of a 2-cm radius steel ball with an initial velocity equal to 1 cm/s. The time histories of the total impact force  $F(t)$ , the impactor position, the beam deflection, the indentation  $\alpha(t)$ , and the in-plane stress at the outer fibre of the beam are given in Nosier *et al* (1992). In Timoshenko's solution the response quantities are obtained up to time  $t \simeq 1778 \times 10^{-6}$  s and therefore only two collisions were observed. The present results, on the other hand, indicate that a third collision also occurs at  $t \simeq 1800 \times 10^{-6}$  s. For this problem, the indentation is small indicating that the contact region is very small. From the numerical results it was concluded that (i) the impact force can be accurately determined by considering a fairly reasonable number of normal modes, and (ii) the accurate determination of stress requires the superposition of more normal modes which can drastically increase the computer execution time. In obtaining the present results, the impact load was represented as a concentrated load (case-I loading model). Also the nonlinear integral equation (43) was solved using the linear Lagrange interpolation functions with a time increment  $\Delta t = 10(\mu\text{s})$ . To check the accuracy of our results, we also solved (43) using the Hermite cubic polynomials. It was observed that when the linear Lagrange functions are used, the solution remains accurate and stable for a wide range of  $\Delta t$  when the impact response of the structure is obtained analytically.

##### Example 2

In this example we consider the impact response of a ten-layered symmetric laminate  $(0^\circ/90^\circ/0^\circ/90^\circ/0^\circ)_s$ , which is also considered by Qian & Swanson (1980), with material properties:  $E_{11} = 120$  GPa,  $E_{22} = 7.9$  GPa,  $G_{12} = G_{13} = 5.5$  GPa,  $\nu_{12} = \nu_{23} = 0.3$ ,  $\rho = 1580$  kg/m<sup>3</sup>,  $a = b = 0.2$  m. The laminate is impacted by a 12.7 mm diameter steel ball ( $\rho = 7960$  kg/m<sup>3</sup>) with an initial velocity  $v_0 = 3$  m/s. It was assumed that the impactor is rigid so that the Hertzian contact coefficient  $K_2$  can be obtained from  $K_2 = (4/3)\sqrt{R_s E_{22}}$ . This way it is found that  $K_2 = 8.394 \times 10^8$  N/m<sup>3/2</sup>. Qian & Swanson (1990) developed numerical results, based on FSDPT with the shear correction factors equal to  $\pi^2/12$ , by using two different techniques. Their first technique was based on the Rayleigh-Ritz method, with numerical integration in time, and the Hertzian contact law. In their second technique they replaced the Hertzian contact law, by a linear contact law

$$F = K_1 \alpha. \quad (69)$$

Therefore, instead of a nonlinear integral equation they analytically solved a linear

integral equation similar to

$$K_1^{-1} F(t) = v_0 t - \frac{1}{m_0} \int_0^t f(\tau) \cdot (t - \tau) d\tau - \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sum_{k=1}^K \bar{B}_{mnk} \int_0^t f(\tau) \cdot \sin \omega_{mnk} (t - \tau) d\tau, \quad (70)$$

by using the Laplace transformation technique. In so doing, they had to find the infinite roots of a polynomial equation by using a numerical scheme. Rather than solving (70) analytically as in Qian & Swanson (1990), if we follow our procedure and use the global linear Lagrange interpolation functions for the approximation of the total contact force  $F(t)$ , we obtain

$$(K_1^{-1} + b_p) \cdot f_p + c_p = 0, \quad (71)$$

which yields

$$f_p = -c_p / (b_p + K_1^{-1}), \quad (72)$$

where  $b_p$  and  $c_p$  are the same as defined earlier. It is our belief that the solution (72) is more direct and less involved than the analytical one proposed by Qian & Swanson (1990). In that paper it was assumed that  $k_1 = 5.866 \times 10^6$  N/m and numerical results based on the Hertzian and the linear contact laws were compared. Three different

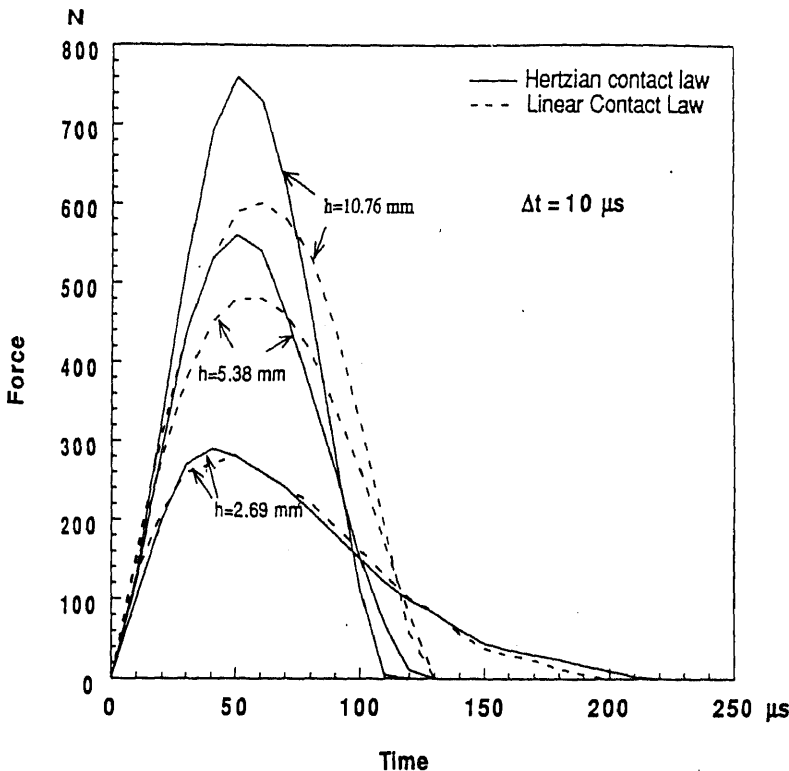


Figure 5. The contact force history of  $(0^\circ/90^\circ/0^\circ/90^\circ/0^\circ)$ , laminate.

thicknesses were assumed for the laminate. It was assumed that the impact load was distributed uniformly over a small square region. To check our analyses and computer codes, we developed the numerical results of figure 5 in that paper which are displayed here in figure 5. The results of Qian & Swanson (1990) are not presented here since they are identical to our results. It is clear from figure 5 that Hertzian and linear contact laws yield almost identical results for a thin plate. This, of course, is due to the fact that the significance of indentation vanishes for thin plates. This conclusion was also reached by Qian & Swanson (1990). It is to be remembered that no clear and scientific methodology yet exists for determination of the contact coefficient  $k_1$  of the linear law.

In order to obtain an assessment concerning the significance of various loading models considered in this report, we have displayed in figure 6 the time history of the impact force of the ten-layered laminate (with  $h = 2.69$  mm). The impact load in figure 6 is modelled according to cases I and VI loading models. Remember that in the case VI loading model the load is assumed to be distributed in accordance to the Hertzian contact law and, furthermore, the time-dependence of the contact region is taken into account. It is seen from figure 6 that the two models yield identical results for a reasonably small  $\Delta t$ . The results of two models are not, of course, identical when a larger time increment is assumed. This is primarily due to the fact that in the numerical evaluation of the second integral appearing in equation (65) we assumed that the contact force remains *constant* during each *small* time increment. Our conclusion here is that since the contact region is often an extremely small region,

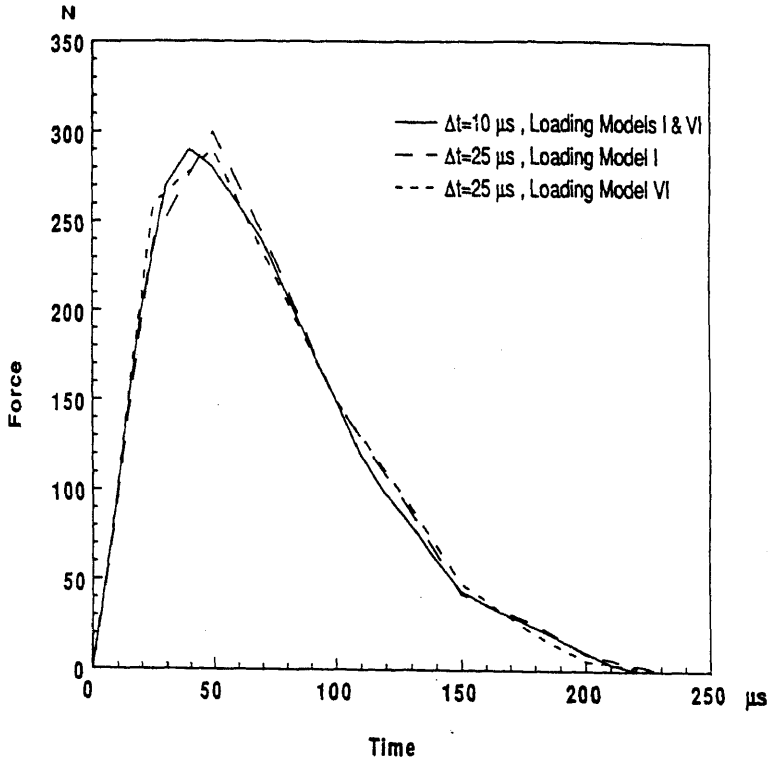


Figure 6. Comparison of contact force predicted by various loading models.

the loading model I will yield very accurate results as far as the total impact force  $F(t)$  and the transverse displacement of the laminate are concerned. This is particularly true when the equivalent single-layer theories are employed, since transverse inextensibility is assumed in such theories. However, it is found, based on the results of LWPT, that a realistic description for both stresses in the vicinity of the contact region and the impact force is only obtained from the case-VI loading model. The calculated values of the total force at various time steps can then be used to obtain the corresponding values of the radius of the circular contact region.

The convergence study of the in-plane stress  $\sigma_{11}$  (see Nosier *et al* 1992 for details regarding the stress calculations), at the point of contact ( $x = a/2, y = b/2, z = -h/2$ ) is accomplished with the help of figure 7. The results indicate that a relatively large number of normal modes are required to achieve convergence for the stress. It will soon become clear that a correct distribution for stresses near the contact region cannot be obtained from any of the single-layer theories during the *contact period*.

When  $h = 2.69$  mm multiple (two) impacts occur in figures 5 and 6, but only the first impact is shown here. This problem was also considered by Sun & Chen (1985) who used a plastic contact law and an experimentally determined value for  $K_2 = 1.413 \times 10^9$  N/m<sup>3/2</sup>. Using this value, and based on FSDPT, we have displayed the time history of the impact force in figures 8 and 9. Comparison of figures 7 and 8 reveals that, as opposed to stress calculation, a relatively low number of normal modes is required for the convergence of the impact force. This conclusion is also valid as far as the transverse displacement is concerned. We should emphasize at this

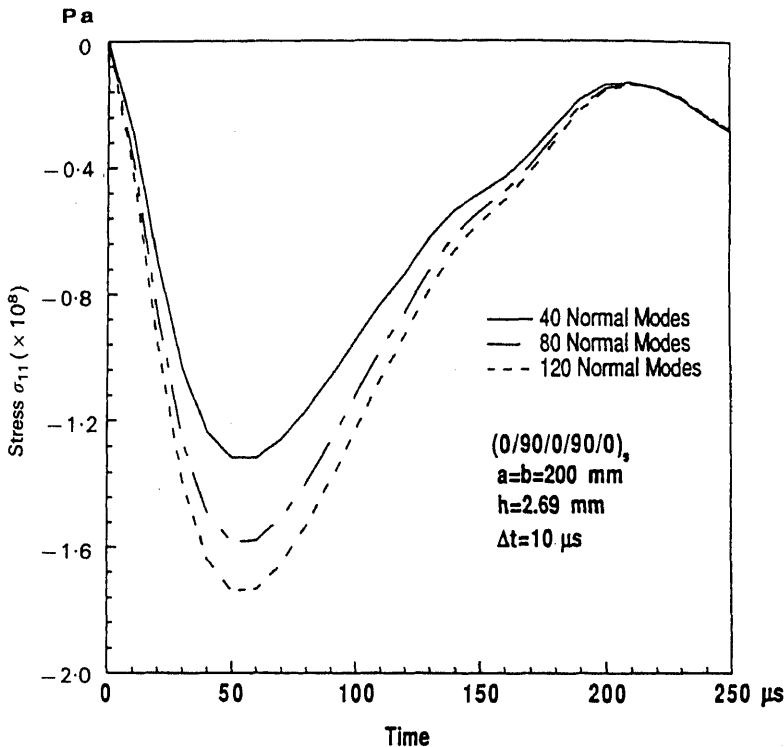


Figure 7. Convergence study for the in-plane stress  $\sigma_{11}$  at the contact point ( $x = a/2, y = b/2, z = -h/2$ ).

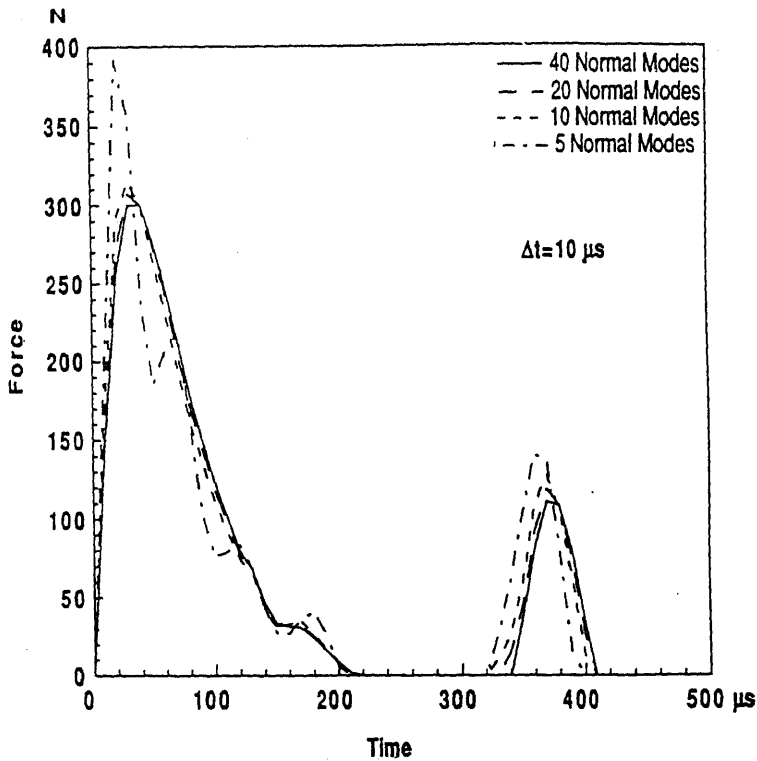


Figure 8. Convergence study for the contact force.

point that this rapid convergence is a direct consequence of the transverse inextensibility (in the thickness direction) assumption in the equivalent single-layer theories. We will demonstrate that the convergence of the impact force and the displacement at the point of contact is quite slow by using LWPT. Furthermore, it will become apparent that none of the response quantities in the vicinity of the contact region is correctly predicted by the equivalent single-layer theories *during the contact period*. The results of figure 9 also indicate that the analytical solution remains stable and accurate for relatively large values of  $\Delta t$ .

### Example 3

Here we consider a three-layered laminate ( $0^\circ/90^\circ/0^\circ$ ) with material properties:  $E_1 = 25.1 \times 10^6$  psi,  $E_2 = 4.8 \times 10^6$  psi,  $E_3 = 0.75 \times 10^6$  psi,  $G_{12} = 1.36 \times 10^6$  psi,  $G_{13} = 1.2 \times 10^6$  psi,  $G_{23} = 0.47 \times 10^6$  psi,  $\nu_{12} = 0.036$ ,  $\nu_{13} = 0.171$ ,  $\rho = 1.4667 \times 10^4$  lb-s<sup>2</sup>/in<sup>4</sup>. We will use the case-VI loading model in the remaining developments. We assume that the square plate ( $a = b = 10$  in.) is impacted by a 0.25 in. radius steel impactor with a mass equal to  $4.8 \times 10^{-5}$  lb-s<sup>2</sup>/in. and an initial velocity of 100 in./s. As we pointed out earlier, the contact coefficient  $K_2$  appearing in the Hertz law must be obtained experimentally, at least for orthotropic laminates. Here by using the modified Hertz law we find that  $K_2 = 2.791 \times 10^6$  lb/in<sup>3/2</sup>. If we follow the procedure suggested by Greszczuk (1982), we obtain  $K_2 = 0.837 \times 10^6$  lb/in<sup>3/2</sup>. In evaluating  $K_2$ , we also assumed that  $\nu_z = \frac{1}{2}(\nu_{13} + \nu_{23})$  and  $G_z = \frac{1}{2}(G_{13} + G_{23})$ . Clearly there exists a big

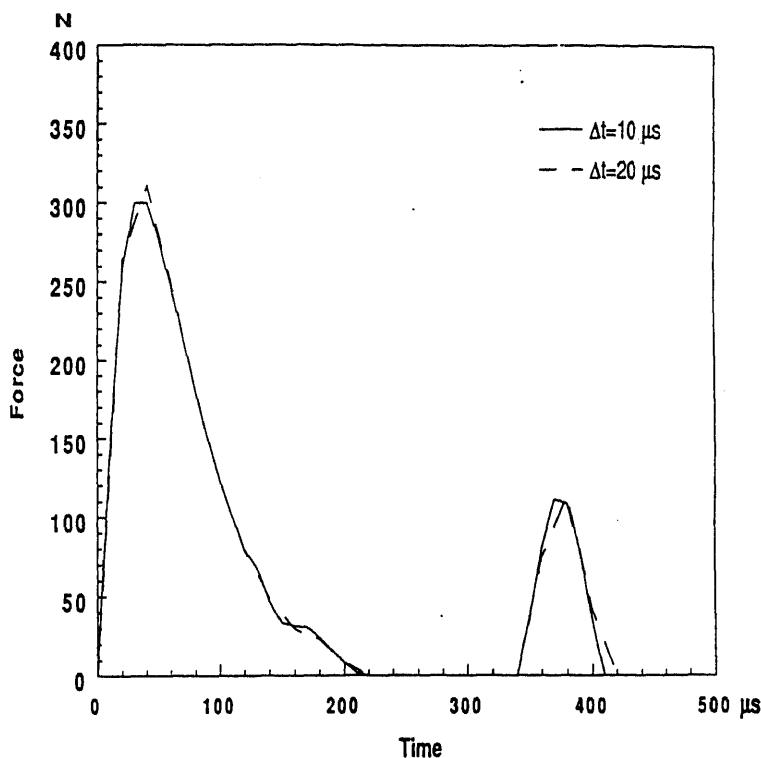
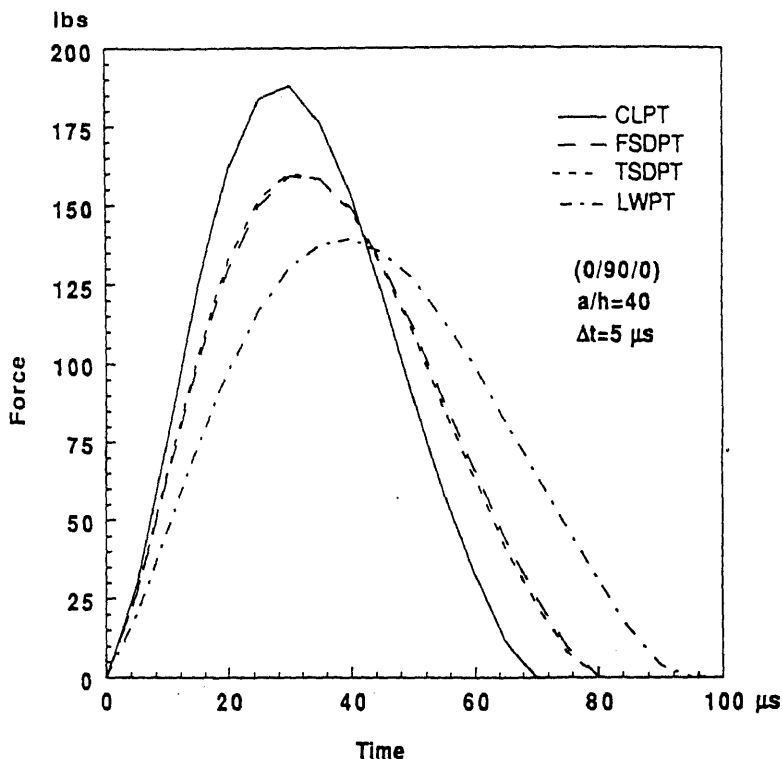


Figure 9. The effect of time increment  $\Delta t$  on the contact force history.

difference between the two values of  $K_2$ . Most likely the actual value of  $K_2$  lies somewhere between these two values. This conclusion is merely based on the observation that the value of the out-of-plane Young's modulus  $E_3$  is much smaller than  $E_2$ . Here, however, for numerical calculations we use the value of  $K_2 = 2.791 \times 10^6 \text{ lb/in}^{3/2}$ . Also, we again use  $\pi^2/12$  for the shear correction factors in FSDPT. We have displayed the time-histories of the impact force and the transverse displacement of the laminate ( $a/h = 40$ ) in figures 10 and 11, respectively. In developing the numerical results based on LWPT we modelled each physical layer as two layers. Figure 10 is clear justification for employing this theory in the impact problems. The inaccuracy of the equivalent single-layer theories in such problems stems from the fact that the assumptions of indentation and transverse inextensibility are made simultaneously. Indeed, from figure 11 it is observed that during contact there is a significant difference between the transverse displacements at the point of contact ( $x = a/2$ ,  $y = b/2$ ,  $z = -h/2$ ) and at the point ( $x = a/2$ ,  $b/2$ ,  $z = h/2$ ). After the contact period the two displacements become essentially identical. The slight differences cannot be distinguished from the figures. For clarity, the time-histories of the displacement of the plate at the contact point, the impactor position, and indentation during contact period are displayed in figure 12.

Similar results are also generated for a thicker plate ( $a/h = 20$ ) and are shown in figures 13 and 14. It is observed that for such a plate there is yet a more significant difference between the transverse displacements at the contact point and at point





**Figure 10.** Comparison of the contact force histories predicted by various theories.

( $x = a/2$ ,  $y = b/2$ ,  $z = h/2$ ). This is the primary reason for the impact force being lower than that predicted by the equivalent single-layer theories.

Results for an additional example, related to impact response of an antisymmetric cross-ply laminate ( $0^\circ/90^\circ$ ) are given in Nosier *et al* (1992).

## 5. Summary and conclusions

In this report we have developed the low-velocity impact dynamic analyses of a laminated plate according to four different plate theories. The first three theories belong to the class of the equivalent single-layer theories in which the transverse extensibility of the plate is ignored. The fourth theory, the one which takes this effect into account by representing the displacements in a layerwise manner, is called the layerwise plate theory of Reddy (LWPT). Throughout this study we have demonstrated the effectiveness of this theory through several numerical examples pertaining to the natural frequencies, the impact force, the displacement, and the in-plane and inter-laminar stress components (Nosier *et al* 1992). In particular, we have found that the rate of convergence of all the response quantities, including the impact force and the transverse displacement, is basically the same and very slow, specially during the contact period and in the vicinity of the contact zone, according to the layerwise theory. More importantly, a full three-dimensional description of the stress field is

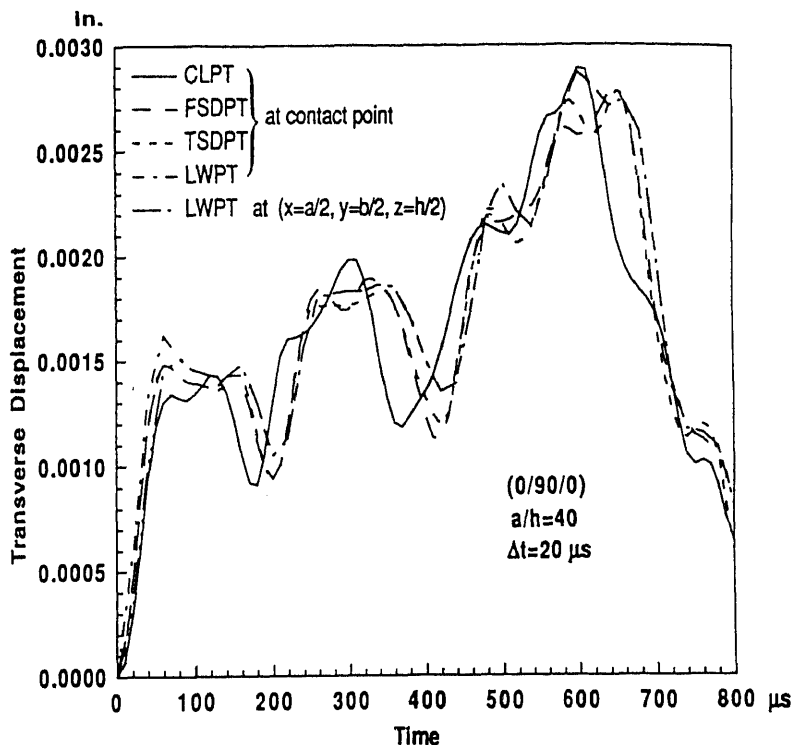


Figure 11. Comparison of the displacement of a laminate at the contact point and at point  $(x = a/2, y = b/2, z = h/2)$ .

obtained only through LWPT. The results of this theory indicate a tremendous amount of stress concentration during the contact period in the vicinity of the contact point. During the contact period, the local deformation in the thickness direction in the neighbourhood of the contact region is the primary reason for the stress build-ups. On the other hand, the nature of the stress is mainly flexural in the free-vibration region. Indeed, for a thick plate the flexural stress is negligible as compared with the infinitely large values of the stress components in the contact region during the duration of the contact. All these observations were made through the employment of the layerwise theory. In fact, only now through the use of LWPT, not through a single-layer-theory, can the matrix cracking, fibre breakage, debonding, and other experimentally observed phenomena be explained. To predict the onset of any failure one should incorporate a failure criterion into our impact analysis. Because of the full 3-D capacity of LWPT, we more realistically modelled the impact pressure in accordance with the Hertzian law of contact. This model, named in this paper as the case-VI loading model, takes into account the time-dependence of the contact area. This, on the other hand, has necessitated the solution of a relatively more complicated nonlinear integral equation. In solving this equation, we made the assumption that the impact force was constant during small time intervals in evaluating the second integral appearing in this equation. In order to be able to use a larger time step in the analysis, the second integral must be, as in case of the first integral, evaluated by assuming a linear variation for the impact force during various time intervals.

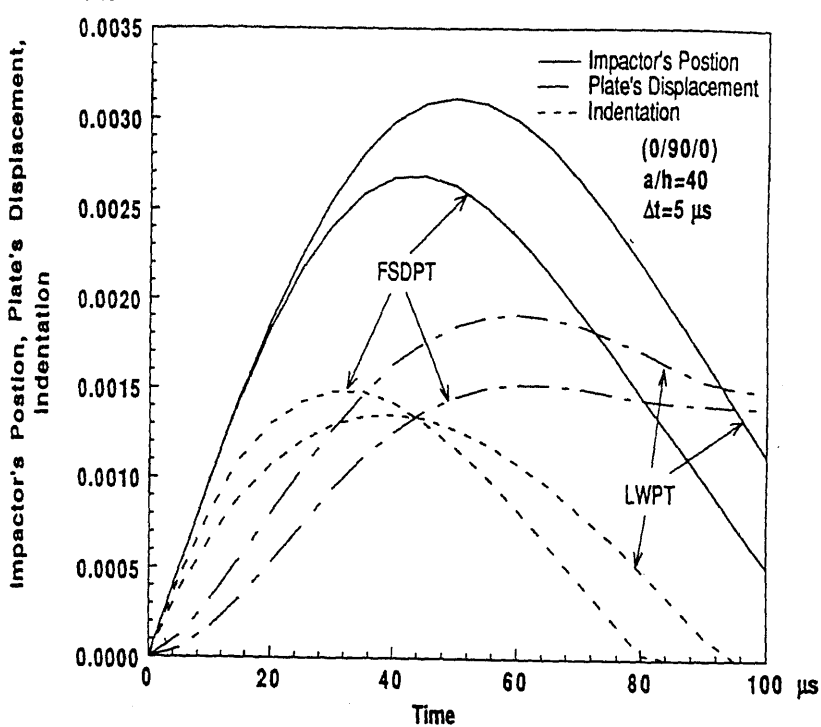


Figure 12. Comparison of the laminate displacement, indentation, and impactor's position at the contact point ( $x = a/2$ ,  $y = b/2$ ,  $z = -h/2$ ).

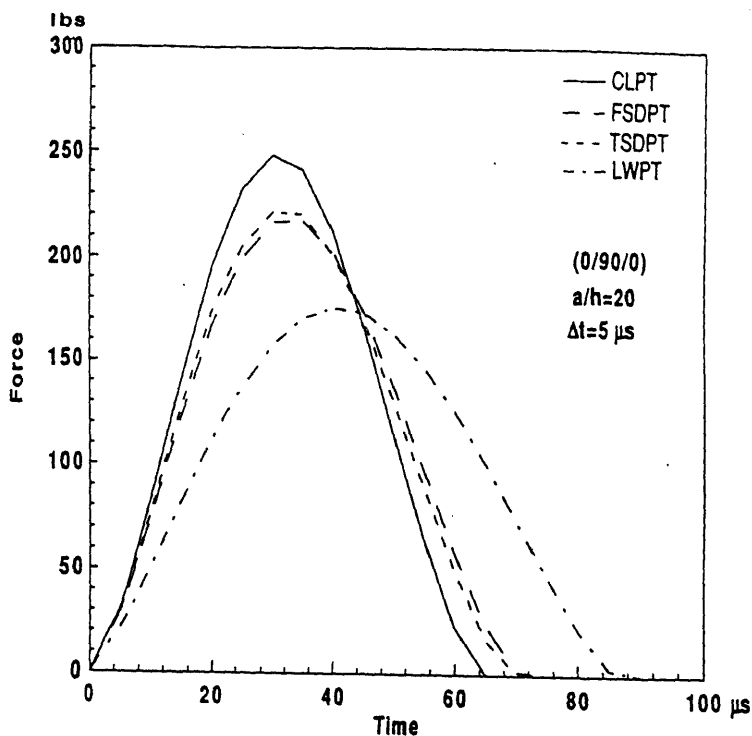


Figure 13. The contact force history of a (0°/90°/0°) laminate ( $a/h = 20$ ) according to various theories.

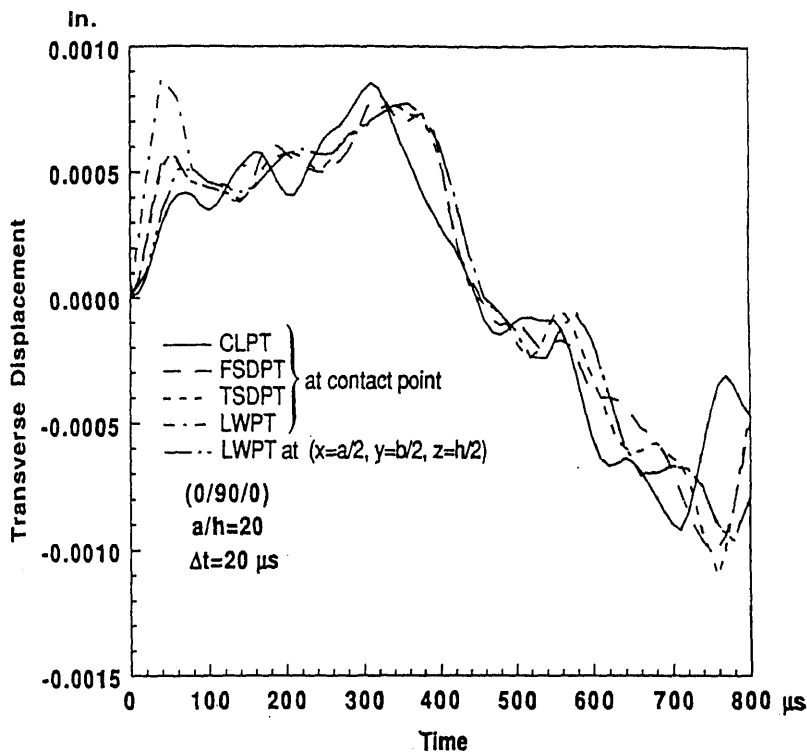


Figure 14. The transverse displacement of  $(0^\circ/90^\circ/0^\circ)$  laminate ( $a/h=20$ ) according to various theories.

As far as the equivalent single-layer theories are concerned, we point out that the response quantities predicted by these theories, especially by the shear-deformation theories, are accurate to a certain degree only in the free-vibration period. During the contact period, the results of these theories are valid only at points away from the contact region. On the other hand, we are often only concerned with a precise description of the stress field in the contact zone neighbourhood because of the existence of high stress concentrations. For this reason the usefulness of the equivalent single-layer theories is questionable.

## References

- Abrate S 1991 Impact on laminated composite materials. *Appl. Mech. Rev.* 44: 155–190
- Bogdanovich A E, Yarve E V 1989 Numerical analysis of laminated plates subjected to impact loading. *Proc. of the American Society for Composites, 4th Technical Conference* (Blacksburg, VA: Virginia Polytechnic Inst.) pp. 399–409
- Bogdanovich A E, Yarve E V 1990 Calculation of damage zones in laminated composite plates subjected to low velocity impact. *Proc. of the American Society for Composites, 5th Technical Conference*, East Lansing, Michigan, pp 1–11
- Byun C, Kapania R K 1991 Prediction of interlaminar stresses in laminated plates using global orthogonal polynomials. *Enhanced analysis techniques for composite materials* (eds) L Schwer, J N Reddy, A Mal, ASME Winter Annual Meeting, Atlanta, pp. 113–124

- Byun C, Kapania R K 1992 Nonlinear impact response of thin imperfect laminated plates using a reduction method. *Compos. Eng., Int. J.* 2: 391-410
- Cairns D S, Lagace P A 1989 Transient response of graphite epoxy and kevlar epoxy laminates subjected to impact. *AIAA J.* 27: 879-884
- Chaudhuri R A, Seide P 1987 An approximation semi-analytical method for prediction of interlaminar shear stresses in arbitrary laminated thick plates. *Comput. Struct.* 25: 627-636
- Chen J K, Sun C T 1985 Dynamic large deflection response of composite laminates subjected to impact. *Compos. Struct.* 4: 59-73
- Conwag H D 1956 The pressure distribution between two elastic bodies in contact. *Z. Angew. Math. Phys.* 7: 460-465
- Goldsmith W 1960 *Impact: The theory and physical behaviour of colliding solids* (London: Edward Arnold)
- Greszczuk L B 1975 Response of isotropic and composite materials in particle impact. *Am. Soc. Test. Mater., Spec. Tech. Publ.* 568: 183-211
- Greszczuk L B 1982 Damage in composite materials due to low velocity impact. In *Impact dynamics* (eds) J A Zukas, T Nicholas, H F Swift, L B Greszczuk, D R Curran (New York: John Wiley)
- Kapania R K, Raciti S 1989 Recent advances in analysis of laminated beams and plates. Part II: Vibrations and wave propagation. *AIAA J.* 27: 935-949
- Kapania R K, Yang T Y 1986 Formulation of an imperfect quadrilateral doubly-curved shell element for post-buckling analysis. *AIAA J.* 24: 310-311
- Kant T, Mallikarjuna 1991 Nonlinear dynamics of laminated plates with a higher-order theory and  $C^0$  finite elements. *Int. J. Nonlinear Mech.* 26: 335-343
- Nosier A, Kapania R K, Reddy J N 1992 Low velocity impact response of laminated plates. Center for Composite Materials and Structures, Report 92-20, VPI & SU, Blacksburg, VA
- Obst A W, Kapania R K 1992 Nonlinear static and transient analysis of laminated beams. *Compos. Eng., Int. J.* 2: 375-390
- Pagano N J 1969 Exact solutions for composite laminates in cylindrical bending. *J. Compos. Mater.* 3: 398-411
- Petersen B R 1985 *Finite element analysis of composite plate impacted by a projectile*. PhD dissertation, University of Florida
- Pintado P, Vogler T J, Morton J 1991 Impact damage development in thick composite laminates. *Compos. Eng.* 1: 195-221
- Qian Y, Swanson S R 1990 Experimental measurements of impact response in carbon/epoxy plates. *AIAA J.* 28: 1069-1074
- Ramkumar R L, Chen P C 1983 Low-velocity impact response of laminated plates. *AIAA J.* 21: 1448-1452
- Reddy J N 1980 A penalty plate bending element for the analysis of laminated anisotropic composite plates. *Int. J. Numer. Methods Eng.* 15: 1187-206
- Reddy J N 1983 Geometrically nonlinear transient analysis of laminated composite plates. *AIAA J.* 21: 621-629
- Reddy J N 1987 A generalization of two-dimensional theories of laminated composite plates. *Commun. Appl. Numer. Methods* 3: 173-180
- Reddy J N 1990 A general third-order nonlinear theory of plates with moderate thickness. *J. Nonlinear Mech.* 25: 677-686
- Sankar B V, Sun C T 1985a Efficient numerical algorithm for transverse impact problems. *Comput. Struct.* 20: 1009-1012
- Sankar B V, Sun C T 1985b Low velocity impact response of laminated beams subjected to initial stresses. *AIAA J.* 23: 1962-1969
- Sun C T 1977 An analytical method for evaluation of impact damage energy of laminated composites. *Am. Soc. Test. Mater., Spec. Tech. Publ.* 617: 427-440
- Sun C T, Chattopadhyay S 1975 Dynamic response of anisotropic laminated plates under initial stress for impact of a mass. *J. Appl. Mech., ASME* 97: 693-698
- Sun C T, Chen J K 1985 On the impact of initially stressed composite laminates. *J. Compos. Mater.* 19: 490-504
- Sun C T, Liou W J 1989 Investigation of laminated composite plates under impact dynamic loading using a three-dimensional hybrid stress finite element method. *Comput. Struct.* 33: 879-884

- Swanson S R, Smith N L, Qian Y 1991 Analytical and experimental strain response in impact of composite cylinders. *Compos. Struct.* 18: 75–108
- Tan T M, Sun C T 1982 Wave propagation in graphite/epoxy laminates due to impact. NASA CR-168057
- Timoshenko S P 1913 Zur Frage Nach der Wirkung Eines Stosses auf Einen Balken. *Z. Math. Phys.* 62: 198–209
- Timoshenko S, Goodier J N 1970 *Theory of elasticity* (New York: McGraw-Hill)
- Thangjitham S, Librescu L, Cederbaum G 1987 Low-velocity impact response of orthotropic plates using a higher-order theory. *Proc. of 28th AIAA/ASME/ASCE/-AHS/ASC Structures, Structural Dynamics and Materials Conference*, pp. 448–457
- Yang S H, Sun C T 1981 Indentation law for composite laminates. NASA CR-65460



## Managing disasters precipitated by natural hazards

V K GAUR

CSIR Centre for Mathematical Modelling and Computer Simulation,  
National Aerospace Laboratories, Belur Campus, Bangalore, 560 037,  
India

**Abstract.** Natural hazards are a piece of the dynamics of the outer working parts of the earth's thermodynamic engines; and their course in many cases cannot be stayed. But disasters caused by them can be significantly minimized by designing and enforcing hazard-resilient land use plans, building codes and other safety and avoidance measures. Current understanding of the space-time characteristics of natural hazards offers considerable insight in designing effective hazard reduction programmes as well as a research agenda that would progressively vitalize this endeavour.

**Keywords.** Natural hazards; disaster reduction.

Life on earth has always been subject to natural catastrophes. Repeatedly, since the dawn of time, sudden or prolonged changes in the supportive environment of the day have wiped out whole species; but evolution proceeded by stimulating the innovative capacities of those surviving, towards the development of evermore viable life-sustaining mechanisms. Human societies too have been hapless victims of recurring scourges wreaked by floods and drought, storms, earthquakes and landslides. The annual toll of life and property exacted by natural hazards has increased rather dramatically since the Industrial Revolution (figure 1). In recent years, for example, between 1964 and 1983, 2.5 million people were killed and over 750 million rendered homeless by disasters caused by natural hazards. A single moderate earthquake ( $M = 6.4$ ) that rocked Maharashtra last September, killed over 15,000 people and rendered a large number destitute. Indeed, with the increasing concentration of human populations and industrial activity in super cities (figure 2), the potential threat of natural disasters the world over continues to grow. This threat is particularly high in the third world where 3 factors combine to expose large unwary communities to grave risk.

First, the survival needs of a burgeoning human family drive an increasing number to occupy ecological niches of marginal energy, unaware of the slow-developing instabilities that might some day turn into disaster. Second, the one-way environmental degradation caused by the continuing loss of shallow aquifers, forest cover, soils and coastal wetlands, mangrove and coral systems, insidiously erodes the very support base of the natural systems on which the majority that subsist near the earth, are so



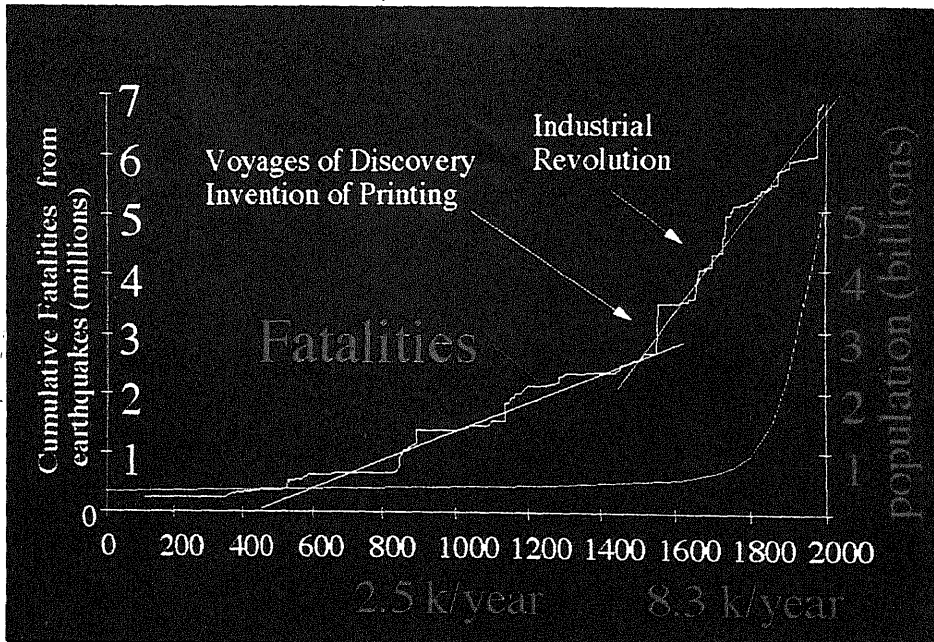


Figure 1. The toll of life exacted by natural hazards.

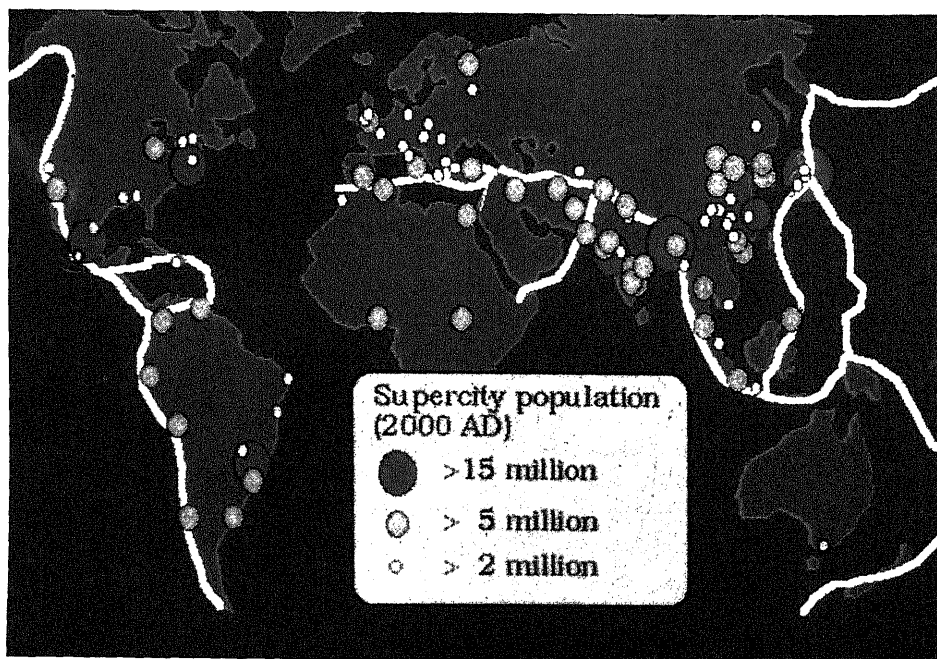


Figure 2. Illustration of the concentration of human population in super cities around the world.

abjectly dependent. Finally, the poor technological infrastructure for generating hard data on critical earth system parameters, encourages a lackadaisical approach to hazard estimation and design of critical structures and preparedness planning strategies. The latter have the further effect of exacerbating the impact of natural hazards by instilling a sense of false confidence in such measures until, as happens so often, they prove quite inadequate when tragedy strikes.

### **Natural hazards – the other face of an energetic planet**

Natural hazards result from extreme fluctuations in the dynamic state of our energetic planet. The earth is a vigorously convecting system powered within by radioactivity and primordial heat, and without by the sun's energy which warms it nonuniformly. Its various spheres of solid, liquid and gaseous domains, while subject to their own time scales of processes, thus constantly flow and exchange materials and energy across their multiple interfaces in a ceaseless struggle to attain equilibrium, in turn, fashioning a wondrous variety of ecological systems in passage. However, it is in the very nature of such variously coupled systems that their meta-stable state would periodically excure to extreme values and even become unpredictably chaotic. Some insights in such processes have been gained in recent years by studying the dynamics of simple coupled systems. Nature, apparently reproduces such behaviour all the time and at various scales. We can visualize these processes at work in the growth of crystals, snow flakes, clouds and drainage channels as well as earthquakes and convective instabilities that breed destructive storms. Natural hazards are thus just the other face of a vibrant youthful planet which alone amongst its four terrestrial sisters organized itself into a system of interactive spheres that would symbiotically regulate its environment for the appearance and sustenance of life forms. Natural hazards are going to be with us forever but disasters need not.

### **Confronting natural hazards for disaster reduction**

The severity of disasters precipitated by natural hazards is, in fact, determined by our ability or otherwise to confront them with knowledge and commitment. Considerable understanding has now been gained of the evolution and characteristic features of cyclones, floods and drought, earthquakes, volcanic eruptions and landslides. Carefully recorded accounts of past events wherever chronicled, have also led to a recognition of their space-time patterns, while analyses of these patterns have created insights in predicting the probabilities of their occurrences in future. Powerful analytical frameworks and computer simulation models have also developed apace with progress in computational techniques, which enable one to translate these prognostications into quantitative descriptions of probabilistic hazards. These figures when incorporated in the design of land-use patterns, hazard-resistant structures and preparedness strategies, can greatly minimize the impact of Nature's fury. For example, by intelligent management of land slopes backed by scientific mapping and monitoring of the area, Japan has been able to considerably reduce the impact of recurrent landslides from 130,000 dwellings destroyed in 1938 to only 2,000 in 1976. In India too, the existence of a network of coastal radars (figure 5 in Joseph, this volume) and satellite-based surveillance systems to track the evolution of cyclonic storms, enabled the authorities to evacuate over 600,000 people to safe shelters in the wake of the 1990 Andhra cyclone, which attained hurricane intensity. The death toll was thus

limited to about 1,000 as compared with 20,000 killed by an earlier Andhra cyclone in 1977.

Satellite-based early warning systems in Bangladesh have similarly helped reduce the fatalities caused by cyclones in recent years: from 300,000 deaths in 1970 to 10,000 in 1985.

These few examples of hazard reduction efforts clearly demonstrate how a focussed application of available technologies can significantly reduce the disaster proportions of a natural hazard. The latter examples, of course, relate only to one aspect of hazard reduction strategy, that is, rescue and relief. Yet more remarkable successes can be achieved by advance planning to ensure long-term protection from natural hazards as well as research to gain a clearer definition of the physical processes which will improve our prediction capabilities. This is shown by the Japanese example founded on systematic assessment of landslide hazard, and design and enforcement of well-formulated building and grading codes.

The recognition of these potential capabilities that now lie within our grasp, in enhancing our resilience to face the rigours of natural hazards, led the United Nations to declare the decade of the Nineties as the International Decade for Natural Disaster Reduction (IDNDR). Serious efforts have accordingly been made in some countries, notably in the West and in Japan, for designing a comprehensive hazard reduction programme and its systematic implementation by exploiting the strong commonalities that Disaster Mitigation measures share, irrespective of the nature of specific hazards.

### **The generic framework for disaster management**

An effective strategy for mitigating disasters caused by natural hazards consists in the design and implementation of four related activities: advance action for long-term protection; preparedness for efficient response to a hazardous event that has happened or is about to happen; recovery and rehabilitation; and research for improving prediction reliability, as well as the design of appropriate engineering and social structures that would effectively embed in the local culture.

#### *Advance planning*

This is the most important activity aimed at providing basic directions for creating an environment for long-term protection and enhancing the resilience of a community in braving natural hazards. The basic grounds prepared in this process also come in handy for implementing rehabilitation measures that would instil confidence in their future viability. This activity involves the following tasks.

- \* Identification of hazard-prone regions on the basis of historical and current knowledge as well as conceptual anticipations.
- \* Evaluation of the probability of exceedance of hazard intensities over various intervals in the future, corresponding to the life spans of different types of structures, utilities and systems; and preparation of hazard intensity maps.
- \* Creation of multihazard Geographical Information Systems and design of land use management schemes.
- \* Design of engineering specifications for various kinds of structures and systems.
- \* Assessment of risks faced by existing structures and systems and designs for retrofitting and relocating them wherever necessary.

### Disaster preparedness

These measures, necessitating horrendous outlays, can obviously be targetted only to specific areas that have already been identified as being vulnerable. They are most effective when well designed protocols for rescue and relief are already in place and can be directed by computer-simulated visualisations of the progress of a suspected or imminent catastrophe. Its basic elements are as follows.

- \* Design and operational readiness of protocols for effective rescue and relief measures, prevention of cascade disasters such as epidemics, and emergency operation of critical services in the event of their failure.
- \* Operational prediction models energized by real-time data for forecasting the progressive evolution of a natural hazard, its estimated space-time characteristics and intensity.
- \* Regular dissemination of information through bulletins carefully designed to evoke a constructive response and avoid panic.
- \* Rapid response action planning and implementation.

### Rehabilitation

The only way to reduce the continuing impact of natural disasters on affected communities is to provide rapid relief and rehabilitation while at the same time freeing them from the spectre of similar catastrophes in the future. In implementing this activity, therefore, *ad hoc* measures should be consciously eschewed and strict adherence enforced to follow the recommended land use pattern and codes for engineering design and control.

### Research and development

While all-out efforts should be immediately launched to use globally available knowledge and technologies in estimating and mapping hazard intensities in threatened areas, and in the development of operational prediction systems within this decade, continual improvement of these capabilities will require an equally focussed effort to close the critical gaps in our understanding, through further research. Nature usually offers considerable revealing information in the wake of highly energetic catastrophes, notably earthquake aftershocks and wind fields after the storm landfall. An important research strategy should therefore aim at meticulous scientific preparedness to glean this information contained in the post-disaster behaviour of earth systems. For example, our understanding of earthquake processes in the Himalaya and in the interior of the continent could have progressed by a quantum jump in the last few years, had Indian scientific organizations been ready to monitor details of the immediate aftershock sequences of the recent Bihar (1988), Uttarkashi (1991) and Khillari (1993) earthquakes, and of changes in the attendant strain field.

An illustrative example to draw attention to some of the outstanding problems in hazard reduction that call for continuing research is given in table 1. <sup>1</sup> are the basic requirements of information products that can be derived from the knowledge already available; and these need to be addressed. The last column spells out the prerequisite state-of-the-art that must be installed to realize these goals. These too, should be installed.

**Table 1.** Some of the immediate problems to be addressed for hazard reduction.

Natural hazard (a)	Outstanding problems for research (b)	Information products based on available knowledge – an immediate concern (c)	Additional state-of-the-art technology systems to make (b) and (c) possible (d)	
Cyclones and storm surges	Sophisticated coupled models for improving the reliability and lead time of forecasts: cyclone tracks, especially their recurvature, time and location of landfall	Improved real-time determinations: Location and velocity of the cyclone eye, wall thickness, wind gradients	Network of Doppler radar systems along the coast and instrumented aircrafts for <i>in-situ</i> determination of storm parameters	
	Heights of storm surges along the coast	Coastal bathymetry data, coastal land topography		
	Wind-structure interactions and design of cost-effective structures and storm shelters	Estimation of extreme wind loads		National facilities for wind tunnel research
	Development of risk assessment frameworks	Estimation of risk to existing structures and systems		
	Earthquakes	Modelling of earthquake source zones: geometry, rates of strain, accumulation		
Rates of strain accumulation and release		Dynamic strain field using repeat GPS monitoring of a few thousand appro- priately located control points	A national facility for continuous monitoring of the crustal deformation fields with subcentimetre accuracy using	

GPS receivers at about 50 control points and repeat yearly measurements at a few thousand control points

Modelling of earthquake processes: initiation and mechanics of rupture

Unfettered availability of high quality seismic records, estimation of ground motion histories, and preparation of hazard intensity maps

Design of regulatory codes for engineering practices and land use management

Geographical information systems; hazard evaluation algorithms

Development of reliable risk assessment frameworks

Estimation of risk to existing structures and systems

Development of improved mapping methods on various scales using satellite remote sensing and ground investigations

Models of landslide processes: initiation and the mechanics of landslide transport, recurrence relations and prediction

Historical records and maps of space-time distribution of landslides; uninterrupted digital data on deformation spanning the clamping pre- and post-landslide regimes

Design of land use patterns and practices and of landslide control measures such as soil drainage and grading codes

Geographical information systems relating to environmental, engineering and cultural parameters; hazard evaluation and maps of hazard intensity.

Development of reliable risk assessment frameworks

Estimation of risk to existing structures and systems

Networks of instrument systems for monitoring the time evolving deformation field in landslide-prone areas

India has taken a number of measures to mitigate the severity of natural disasters caused by floods, storms and drought, but a serious programme to reduce their impact to the levels now possible, through a systematic and focussed application of science and technology, still remains a distant goal. A basic impediment to achieving this goal is the absence of an authentic and reliable information system for environmental data, which forms the core of a disaster-reduction programme.

The extant structure of the government-run scientific organizations charged with the responsibility of data generation, archiving and dissemination, appears to have some inherent difficulties in developing state-of-the-art technology systems capable of generating high quality uninterrupted data sets needed for modern analysis. This fact has to be confronted honestly and squarely whilst establishing the basic information systems. It may even be helpful to seek some insights into the sociology of scientific data-gathering organisations in the country from professionals in the field, so that our next steps in this direction prove to be as vital and flawless as anywhere in the world.

The papers contained in this volume were presented at a short symposium on Disaster Reduction organised at the University of Roorkee as a part of the last annual meeting of the Indian Academy of Sciences. They describe in detail the nature of threats posed by various kinds of natural hazards, and our current endeavours to mitigate their impact.

## Tropical cyclone hazards and warning and disaster mitigation systems in India

PORATHUR V JOSEPH

Satellite Township, Kakanad, Cochin 682 030, India

**Abstract.** The main features of a cyclone that cause death and destruction are: (1) Storm surge, a rapid increase in sea level along the coast, primarily caused by the strong surface wind field of the cyclone as it approaches the coast, (2) the violent sustained wind and wind gusts and cyclone-spawned tornadoes, and (3) the heavy rain and consequent flooding. The paper describes the structure of a tropical cyclone of hurricane intensity and its damage potential, cyclone detection, tracking, forecasting and warning systems. It concludes that with over a hundred years of research and operational experience on cyclones available with the meteorological community and with the INSAT and cyclone detection radar network, India has now an efficient cyclone warning system. This system would however benefit from further sophistication notably the use of Doppler Radars for cyclone wind field monitoring and an aircraft reconnaissance facility for probing cyclones. Research is needed to develop techniques for better forecasting of the tracks of cyclones more than a day ahead and also their intensity changes.

Cyclone disaster mitigation arrangements were organised in the maritime states of India only during the last 25 years. This relatively young field requires more developmental work, particularly in coastal area planning to reduce property losses, and exploration of alternative approaches to large-scale evacuation of the coastal population threatened by cyclones to distant temporary shelters.

**Keywords.** Hurricane intensity; *T*-number; eye and wall cloud; cyclone alert and warnings; cyclone disaster (distress) mitigation.

### 1. Introduction

It was Henry Piddington, President of the Marine Courts, Calcutta who pioneered scientific studies on tropical cyclones in the Indian seas, systematically collecting meteorological logs of ships plying in those waters. He published a series of memoirs in the *Journal of the Asiatic Society of Bengal* during 1838–1858 dealing with individual cyclones. He also wrote a book entitled *The Sailor's Horn-Book for the Law of Storms*, the fourth edition of which appeared in 1864.

One of the oldest recorded cyclonic storms that caused heavy casualties in India



was the one that hit the mouth of the Ganges near Calcutta on October 7, 1737. It is reported to have killed 3,00,000 people and destroyed 20,000 crafts of various descriptions, although there are some doubts about these figures. The river rose by 40 feet over its usual level. These details are taken from a catalogue of 112 recorded cyclones in the Bay of Bengal, up to the end of 1876, compiled by Henry F Blanford, Meteorological Reporter to the Government of India and published in the *Journal of the Asiatic Society of Bengal*. In recent times, the cyclone that crossed the Andhra coast on 19 November 1977 took a toll of 10,000 lives, mainly caused by the 5-metre high storm surge that accompanied it.

Following a devastating cyclone that struck Calcutta in October 1864, the then Government of India established a cyclone-warning system for the port of Calcutta. The India Meteorological Department was born in 1875 and cyclone warning for ports, coastal areas and ships in the Indian seas became one of its routine activities. In 1969 the Government of India advised the governments of all maritime states to set up 'Cyclone Distress Mitigation Committees' (CDMC) in their respective states to effect suitable measures for mitigating the attendant hazards. An efficient, full-fledged system of cyclone 'alerts' and 'warning' has since developed in the country to cope with this recurrent natural threat.

A Tropical Cyclone Programme (TCP) was established by the World Meteorological Organisation (WMO) in 1971 to assist vulnerable countries in minimising the loss of life and property caused by these storms. For disaster prevention and preparedness measures, the TCP takes cooperative action with the office of the United Nations Disaster Relief Coordinator (UNDRO) and the League of Red Cross and Red Crescent Societies (LRCS). Five different bodies look after the regional components of the TCP. One such is the WMO/ESCAP Panel on Tropical Cyclones serving the countries bordering the Arabian Sea, the Bay of Bengal and the northern Indian Ocean. They are: Bangladesh, India, Maldives, Myanmar, Pakistan, Sri Lanka and Thailand. This panel coordinates the cooperative activities among the member countries in tackling cyclone problems. Under a Cyclone Operational Plan for the region, New Delhi issues 'Tropical Weather Outlook' daily at 0600 UTC and 'Cyclone Advisory Bulletins' at 6-hour intervals to the member countries, whenever a cyclonic storm develops in the northern Indian Ocean.

## 2. Tropical cyclone and the associated hazards

A typical mature tropical cyclone is a warm core vortex in the atmosphere (anti-clockwise vortex rotation in the northern hemisphere and clockwise in the southern), cyclonic in the lower troposphere and anti-cyclonic in the upper troposphere. The circulation extends horizontally to some 1000 kilometres from the centre and vertically to about 15 km above sea level. There is an 'eye' at the centre of the cyclone of radius 5 to 50 kilometres. The eye is rain-free with light winds. It is surrounded by a 'wall cloud' made up of tall cumulo-nimbus clouds rising upto an altitude of 15–18 km, the wall cloud thickness being about 10–15 km radially. Below the wall cloud are found the strongest surface winds of the cyclone ( $V_{\max}$ ) and its heaviest rain intensity. Beyond the wall cloud, surface wind speeds decrease gradually with the radial distance from the centre and rainfall is confined to the regions covered by the inward spiralling cloud-bands (composed of cumulo-nimbus clouds and some cumulus clouds at large

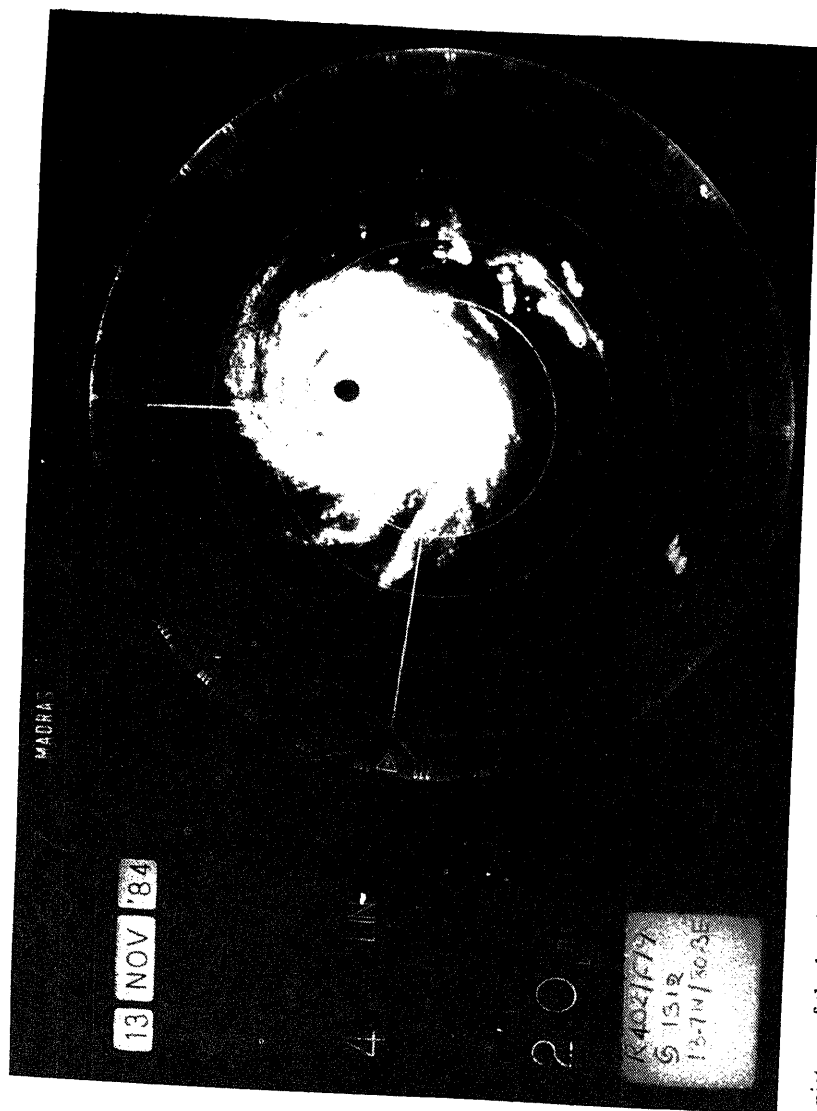
distances from the centre) that are seen within a radial distance of about 400 km from the centre of the cyclone.

As one moves from the periphery to the centre of the cyclone, the sea level atmospheric pressure falls continuously, the largest radial pressure gradients of 2–4 hPa/km occurring in the wall cloud region. Under the influence of frictional forces, the low-level wind direction which is almost tangential to the nearly circular isobars of the cyclone field at about 1 km above sea level, swerves to about 25° towards low pressure at sea level. The low level winds rich in moisture thus possess strong tangential and radial components causing the air parcels to spiral inwards from the peripheral regions of the cyclone towards its centre. In consequence, their rotational velocity (tangential wind) increases rapidly due to partial conservation of its angular momentum. The radial component of the wind converges large amounts of moisture to the central regions of the cyclone, which ascends and condenses in cloud formations there, keeping the central regions warmer than the surrounding tropical atmosphere. This warm anomaly which reaches a maximum of about 15°C at 300–200 hPa level (9–12 km altitude) reduces the radial pressure gradients at these high altitudes. The cyclonically rotating air parcels rising up in the central regions of the cyclone move outwards in the upper troposphere, under the action of unbalanced centrifugal forces (with reduced pressure gradients) and conserving angular momentum, begin to reverse their cyclonic rotation as they move further away from the centre. Satellite pictures of tropical cyclones indeed show both the inward spiralling low-level clouds and the outward-moving cirrus clouds at the upper levels.

Our current knowledge of the structure of tropical cyclones has come from studies made for over a hundred years of different cyclone-prone regions of the world. Early studies using ship reports and measurements from coastal and island observatories gave a reasonably good picture of the surface level features of the cyclone, but that of their three-dimensional structure has been derived mainly from 'reconnaissance flights' using specially equipped aircraft that were flown into the cyclone at various levels, measuring winds, temperatures, humidity and pressure. Direct sensing as well as remote sensing methods, especially compositing of data from routine balloon soundings of the atmosphere has also yielded a wealth of information (Gray 1979, pp 155–218). Details regarding the eye, wall cloud and the spiral cloud bands were obtained from satellite pictures (polar orbiting and geo-stationary) as well as cyclone detection radar systems installed at coastal and island stations. Figure 1 shows the cloud picture of a hurricane intensity tropical cyclone recorded by the cyclone detection radar at Madras.

Low pressure systems such as the tropical cyclones that occur over the oceans are classified in intensity or strength according to the highest sustained surface wind speeds (averaged over a few minutes) and not the instantaneous gust speeds of their circulations. The classification used in India is given in table 1. According to this, a low pressure vortex with winds of 64 knots (32 m/s or 115 km/h) and more is called a severe cyclonic storm with a core of hurricane winds. Similar systems are called 'hurricanes' in the Atlantic and the eastern Pacific and 'typhoons' in the western Pacific Ocean.

By comparing characteristics of the central cloud areas and the eye and spiral cloud bands of a cyclone abstracted from satellite cloud pictures with reconnaissance aircraft measurements of  $V_{\max}$ , Dvorak (1975) suggested a method of estimating the intensity of a tropical cyclone on a T-number scale of 1 to 8. Each T-number has a



cloud picture of the hurricane intensity tropical cyclone recorded by the cyclone detection radar at Madras at 1207 UTC (1737 November 1984). The eye, wall cloud and the closely wound spiral cloud bands of the cyclone (like a clock spring) may be seen. Marker circles are at 40-kilometre intervals, with Madras at the centre. This cyclone crossed the coast a little north on the following day at about 0830 IST. At the time this picture was taken, its intensity was T-6.

**Table 1.** Classification of low pressure systems in India.

Type of low pressure system	Highest sustained surface wind (speed, knots) in the circulation
Low	Less than 17
Depression	17 to 27
Deep depression	28 to 33
Cyclonic storm	34 to 47
Severe cyclonic storm	48 to 63
Severe cyclonic storm with a core of hurricane winds	64 and above

corresponding  $V_{\max}$ . About 70% of the cyclones studied by Dvorak (1975) in both the Atlantic and Pacific Oceans have been shown to grow in intensity at the rate of one T-number per day.

In their mature stages, tropical cyclones the world over show great similarities in their structure and organisation. These similarities result from the strong physical forces associated with the release of latent heat and the relatively uniform characteristics of the lower boundary, i.e. the sea surface (Anthes 1982). The T-number scale developed from studies of cyclones elsewhere in the tropics, is also used in this country for determining the intensity of tropical cyclones in the Indian seas, although it would be desirable to design a T-scale specific to Indian cyclones by combining cloud picture information derived from INSAT with data systematically generated from reconnaissance flights. For, although Indian Ocean cyclones are similar in many respects to hurricanes and typhoons, they are smaller in horizontal extent than their counterparts in the Atlantic and the Pacific.

$V_{\max}$  (in knots) of a tropical cyclone has been found to be related to the pressure drop (PD) in hPa (the difference between the pressures at the periphery and the centre of a cyclone) by a relation of the form  $V_{\max} = k(\text{PD})^{1/2}$ , where  $k$  is a constant for a given ocean basin. For cyclones of the Indian seas, Mishra & Gupta (1976) have derived the value of  $k$  to be 14.2. The relationship between T-number,  $V_{\max}$  and PD applicable to the cyclones of the Indian seas (after Dvorak 1975 and Mishra & Gupta 1976) is given in table 2. In Indian seas T-7 is the maximum intensity so far observed, the corresponding  $V_{\max}$  being about 140 knots. The Andhra cyclone of November 1977, for example reached an intensity of T-7 on maturity on November 18, and maintained that strength till it crossed the coast on the following day. Higher intensities of T-8 have been attained by cyclones in the Pacific and the Atlantic Oceans.

During the 100-year period 1891–1990, there have been 561 cyclones in the Indian seas with maximum winds of 34 knots or more. About 23% of these, according to the India Meteorological Department, intensified into hurricane strength attaining maximum winds of 64 knots or more, the corresponding T-number being 4 and PD equal to 21 hPa (table 2). Of a sample of 181 cyclones which crossed the Indian coast north of 15°N during the 90-year period 1891–1980 analysed by the author and coworkers (Joseph *et al* 1981) only 19 (10.5%) attained an intensity of T-5 ( $V_{\max}$  90 knots and PD 40 hPa) and two reached intensities of T-6 and T-7 respectively.

Cyclones over the Indian seas generally form between latitudes 5°N and 18°N during the pre-monsoon (April, May and early part of June) and post-monsoon (late September to December) seasons. During the monsoon period June to September,

**Table 2.** Relationship between T-number,  $V_{\max}$  and pressure drop of a tropical cyclone in the Indian Seas (after Dvorak 1975 and Mishra & Gupta 1976).

T-number	Maximum wind $V_{\max}$ (knots)	Pressure drop PD (hPa)
1	25	—
2	30	—
2.5	35	6
3	45	10
3.5	55	15
4	65	21
4.5	77	29
5	90	40
5.5	102	52
6	115	66
6.5	127	80
7	140	97
7.5	155	119
8	170	143

weaker systems (mostly depressions) form over the northern part of the Bay of Bengal (north of  $18^{\circ}\text{N}$ ) and then move West-North-Westwards across North India shedding copious rainfall along their paths. This is also the period when a large number of intense hurricanes and typhoons form over the northern hemisphere Atlantic and Pacific Oceans. The occurrence of land to the north of the Indian Ocean and the large vertical shear of the wind in the monsoon atmosphere prevents intensification of these depressions into intense cyclones.

Some of the low pressure areas forming over the warm waters of the Arabian Sea and the Bay of Bengal slowly deepen into depressions. Those of them which further intensify become cyclonic storms within about two days. Some of these further intensify and develop an eye over the next few days, surrounded by a wall cloud and tightly wound spiral cloud bands with windspeeds  $V_{\max}$  of 64 knots or higher and hurricane force winds extending to 50 km or even 100 km from the centre. As the Arabian Sea and the Bay of Bengal are comparatively small in area and bounded by land on three sides, many Indian cyclones encounter the coast (land-fall) early enough and thereby weaken even whilst they are in the stage of intensification. For, tropical cyclones weaken rapidly after their centre (the eye) crosses the coast primarily due to the absence of the energy laden moisture source below. This partly accounts for the low number of T-6 and T-7 cyclones in the Indian seas.

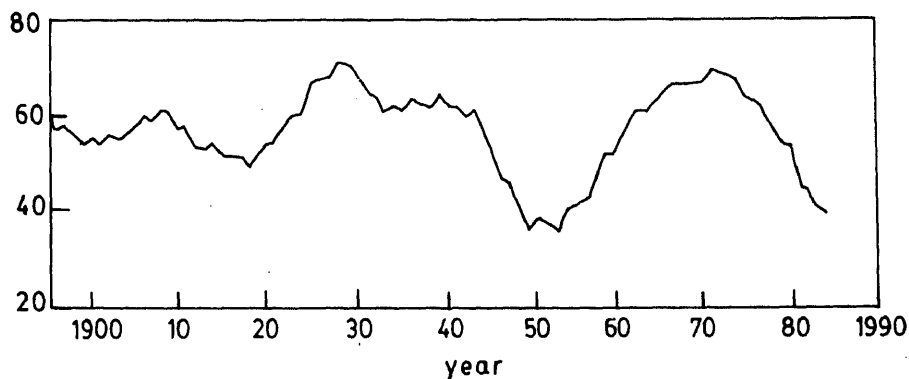
Decadal frequency of tropical cyclones ( $V_{\max}$  34 knots or more) in the Arabian Sea and the Bay of Bengal during the 100-year period 1891–1990 is shown in table 3. The moving ten-year totals of these storms are given in figure 2. Cyclone frequency (decadal) had a minimum around 1950; and maxima around 1928 and 1970. Currently we seem to be passing through an epoch of minimum cyclone frequency. It is notable that the observed decadal frequency variation between the late 1960s and the late 1980s by a factor of two belongs to the era of satellite monitoring, when no cyclone was missed. What then could be the cause of this large variation in cyclone frequency? And will an era of high cyclone frequency return in the future and, if so, when?

**Table 3.** Decadal frequency of tropical cyclones in the Arabian Sea and the Bay of Bengal (data provided by Dr G S Mandal).

Decade	Number of cyclonic storms ( $V_{\max}$ 34 knots and more)
1891–1900	58
1901–1910	58
1911–1920	52
1921–1930	67
1931–1940	61
1941–1950	51
1951–1960	41
1961–1970	63
1971–1980	64
1981–1990	38

Studies at the Colorado State University on the spatial distribution of the frequency of tropical cyclones during the 20-year period 1958–1977 (Gray 1979) have shown that the climatological cyclone genesis depends on several parameters such as sea surface temperature (SST) and the depth of warm water (ocean thermal energy above the level of 26°C in the top 60 metres of the ocean), the vertical shear between horizontal winds at 950 and 200 hPa, middle tropospheric relative humidity, cyclonic vorticity in the low level air flow and the degree of convective instability in the atmosphere that is necessary for the formation of cumulo-nimbus clouds without which tropical cyclones cannot develop. Some of these parameters may be responsible for the observed decadal variability. An indepth study of this problem may create insights for forecasting active and weak epochs of cyclone frequency, several years in advance. Landsea (1993) has reported that intense hurricanes over the Atlantic Ocean during the 1970s and 1980s were much less frequent than those experienced during the 1950s and 1960s.

On an average, about 80 tropical cyclones with sustained winds of 20–25 m/s (40–50 knots) occur each year in the seven cyclone prone ocean basins (Gray 1979).



**Figure 2.** Ten-year moving totals of number of cyclonic storms in the Arabian Sea and the Bay of Bengal with  $V_{\max} = 34$  knots or more during the period 1891–1989. The figure was provided by Dr G S Mandal.

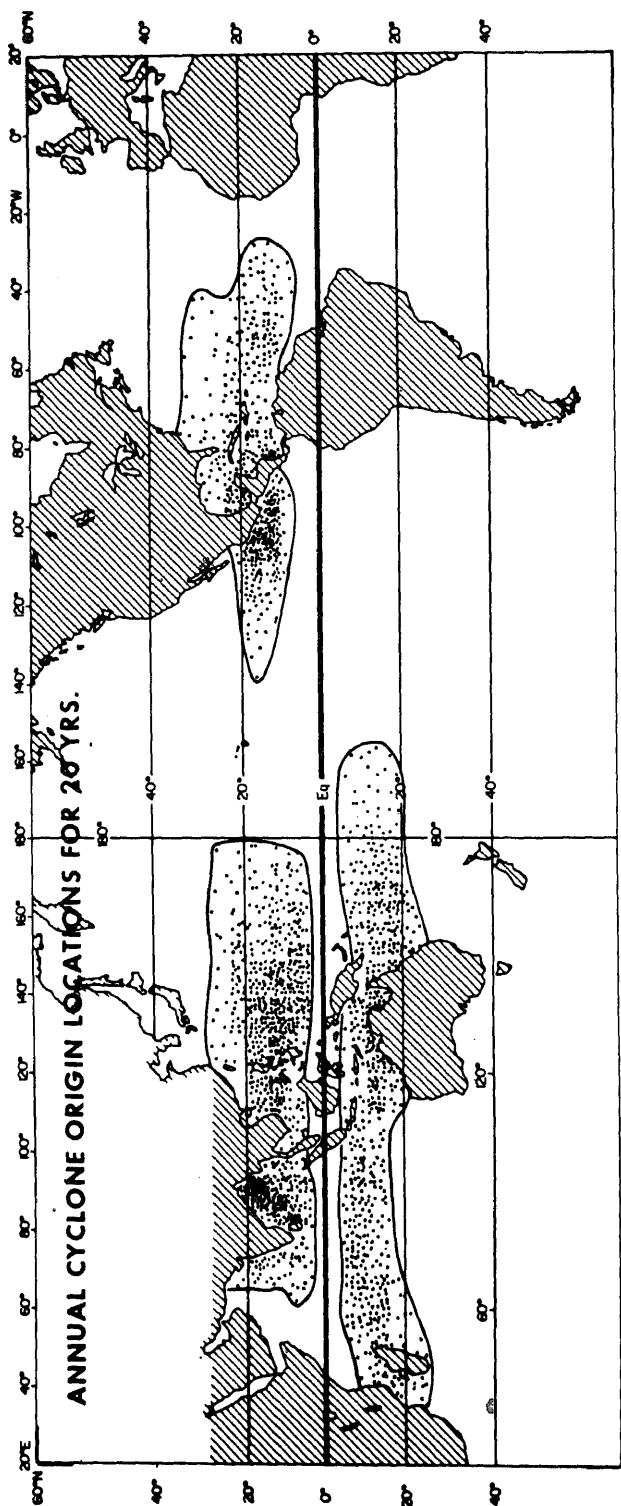


Figure 3. Location of the genesis points of tropical cyclones over a 20-year period 1958 to 1977 (Gray 1979, p. 160)

**Table 4.** Annual frequency of tropical cyclones by ocean basin-average of 20 years 1958 to 1977 (Gray 1979, pp. 155–218).

North Indian	Northwest Pacific	Northeast Pacific	Northwest Atlantic	South Indian	Australian	Southwest Pacific	Total
6.4	26.3	13.4	8.8	8.4	10.3	5.9	79.5

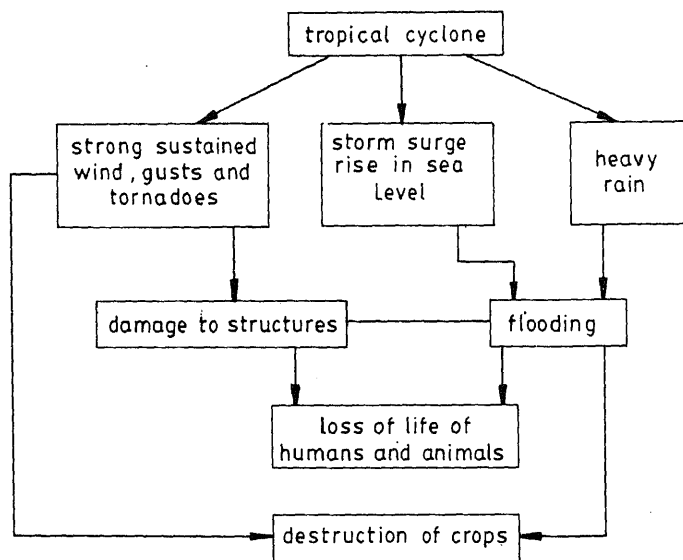
Figure 3 shows the location of the genesis points of these cyclones for the 20-year period 1958–1977; the ocean basin-wise frequency is given in table 4. Tropical cyclones are completely absent in the cold ocean waters of the South Atlantic and the eastern South Pacific, nor do they form within 4–5° latitude of the equator. The western Arabian Sea does not produce tropical cyclones either. Nearly 65% of the global tropical cyclones form in the zone bounded by 10° and 20° latitude.

In a survey of the period 1964–1978 (Southern 1979), it was found that cyclones caused maximum deaths when compared to other natural disasters (table 5). In November 1970 a single cyclone took a toll of 3,00,000 lives in Bangladesh. Over the years, as methods for forecasting the track and intensity of cyclones have improved, disaster mitigation efforts have made remarkable progress, greatly reducing the death toll, although damage to property has steadily increased due to unplanned expansion of human habitats and activities in coastal areas subject to cyclone threats. Figure 4 shows the potential major impacts of a tropical cyclone upon landfall. The important factors are: (i) storm surge, that is, a sudden rise in sea level along the coast, primarily caused by strong surface winds of the cyclone as it approaches the coast, (ii) strong sustained wind, wind gusts and tornadoes and (iii) heavy rain which in many cases is of the order of 30–50 centimetres per day over large areas. In addition to the loss of human and animal lives, damage to structures and destruction of crops as shown in the figure, there are attendant dislocations like loss of communication and power,

**Table 5.** Natural disasters (1964–1978).

Disaster	Total deaths in 14 years	Greatest single event
Tropical cyclones	416 972	300 000 (Bangladesh, 1970)
Earthquakes and tidal waves	195 328	66 794 (Peru, 1969)
Floods	26 724	8 000 (S. Vietnam, 1964)
Tornadoes, severe local storms	4 062	540 (Bangladesh, 1969)
Avalanches and landslides	5 790	1 450 (Peru, 1974)
Volcanic eruptions	2 572	2 000 (Zaire, 1973)
Extratropical cyclones	1 860	166 (USA, 1966)
Heat (cold) waves	505	291 (India, 1973)





**Figure 4.** The potential major impacts of a tropical cyclone upon crossing the coast (landfall).

the erosion of beaches, damage to off-shore installations, damage to shipping and fishing, loss of soil fertility due to saline incursion, and contamination of domestic water supply (Anthes 1982; Southern 1987).

### 3. The cyclone warning system

The India Meteorological Department is the agency responsible for monitoring cyclone development in the Indian seas and for the issue of timely alerts and warnings regarding their temporal evolution and associated weather hazards. It functions through the Area Cyclone Warning Centres (ACWC) at Bombay, Calcutta and Madras, and the Cyclone Warning Centres (CWC) at Ahmedabad, Bhubaneshwar and Vishakhapatnam. With the routine availability of hourly cloud pictures now, taken by the Indian geostationary satellite INSAT (visible during day light hours and infra-red round the clock) the problem of timely detection of tropical cyclones that generally form far out in the sea, and of following their tracks has been largely solved. Ten high power cyclone detection radars along the east and west coasts of India are capable of seeing the eye, wall cloud and the spiral cloud bands of a cyclone out to 400 km away from the coast (about the distance covered by a cyclone in a day). The Cyclone Detection Radar (CDR) network of India is shown in figure 5. Adjacent radar stations have overlapping range circles so that the same cyclone can be seen by more than one station at any time when they are close to the coast. With INSAT and the CDR we now have a dependable system for cyclone warning. Radar systems and satellites employ remote sensing techniques. It has been reported however that errors in determining the cyclone centre can be up to 110 km by satellite fixes, 20–55 km by radar observations and about 20 km by aircraft reconnaissance (Elsberry 1987, pp. 1–12, 91–131). Additional data required come from meteorological measurements

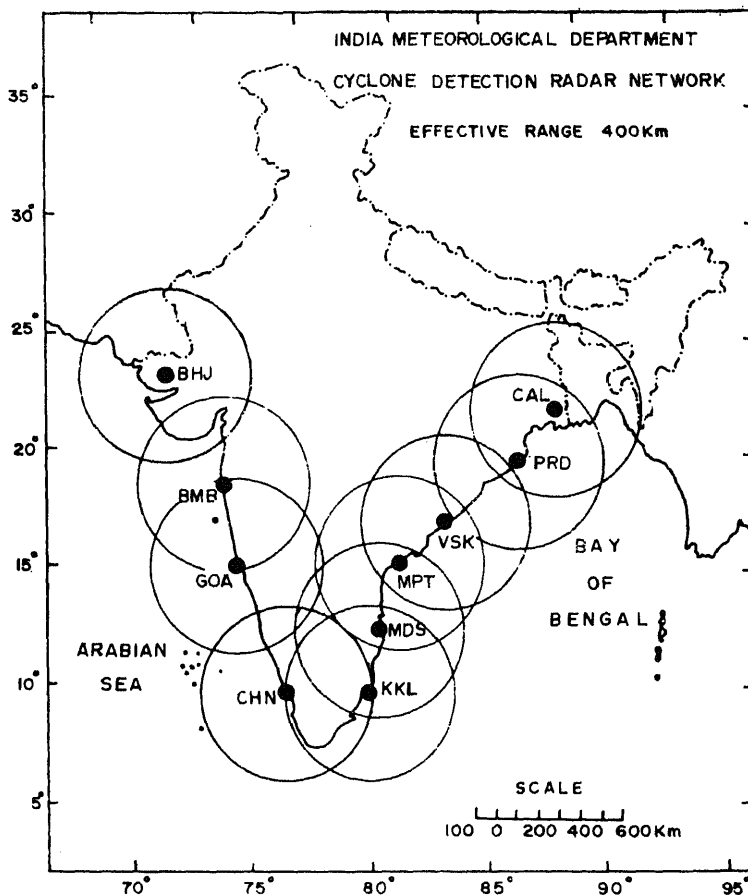


Figure 5. The cyclone radar network along the Indian coast line.

taken by coastal and island observatories and from ships in the field of the cyclone whose data routinely flow into the cyclone warning centres. What we do not have is a reconnaissance aircraft facility. These specially instrumented aircraft provide accurate information about the location of the cyclone centre and other details regarding its intensity and the spatial distribution of winds. Such a facility used to be provided by the Department of Defence (DOD) of USA for the collection of data from hurricanes in the Atlantic and typhoons in the Pacific, but was discontinued for the North-West Pacific in August 1987. Gray *et al* (1991) studied the technical aspects of this issue at the instance of the American Meteorological Society and found that although weather satellites are absolutely essential for tropical cyclone observation, satellite measurements of position, intensity, ambient wind distribution etc. of a cyclone are sometimes degraded leading to inaccuracy in forecasts. Additional data from reconnaissance flights can offset this drawback.

Doppler radars have capabilities to provide details of the cyclone's wind field in addition to usual data on the eye, wall cloud and spiral cloud bands produced by conventional radars. Their ability to continuously monitor the wind maxima associated with the wall cloud and the radial extent of the hurricane strength wind field can be

of great assistance to the forecaster. The United States has planned to deploy Doppler radars round the country during the 1990s (WMO 1990). The India Meteorological Department too, it is understood, has plans to replace the existing CDR network with Doppler radars in a phased manner.

The most important element to be forecast is the future path or track of a cyclone. Where will it cross the coast is also very important. Many of the Indian cyclones move Westwards and North-Westwards. Their tracks are easier to forecast. Others while moving initially in these directions, later turn Northwards. Some of them turn East of North; these are the 'recurving' cyclones and forecasting their track is a difficult problem, often involving large errors.

Objective schemes have been developed for forecasting cyclone tracks, based on persistence, climatology and steering. A persistence forecast uses a linear extrapolation of the smoothed observed track of the cyclone during the previous 12 or 24 hours. The climatological forecast moves the cyclone in the resultant direction with the average speed of movement of all the past cyclones over a 50 to 100-year period in a small area around its present location and a small calendar time interval (say one or two weeks) on either side of that day. An objective method combining both these approaches called CLIPER is in wide use today. India also uses this scheme in operational cyclone warning and advisories are issued to forecasters from a central office. The large scale atmospheric flow in which the tropical cyclone is embedded is empirically related to the cyclone track. This phenomenon is called 'steering', and is derived routinely from the twice-daily balloon soundings of the atmosphere from a large network of radiosonde stations world-wide and from cloud motion vectors obtained from geo-stationary satellites like INSAT. The steering concept is very useful in cyclone-track prediction, particularly to forecast recurvature.

Forecasters can also get guidance from statistical-dynamical models in which the cyclone motion is statistically related to the large scale atmospheric flow features, actually observed or forecast. For cyclone forecasting beyond a day, one also needs an accurate prediction of changes in the large scale circulation around the cyclone. The National Centre for Medium Range Weather Forecast (NCMWRP) established at New Delhi with a super-computer facility can greatly empower the cyclone forecaster in this endeavour by forecasting the evolution of the large scale atmospheric flow around India a few days ahead.

Forecasting intensity changes of cyclones is another area where empiricism prevails. While cyclones continue to intensify (many at one T-number per day) over the warm seas, and reach maturity at some T-number, what may be the environmental factors in the atmosphere and oceans that determine the maximum intensity of a cyclone and its growth rate prior to reaching that intensity? This is an area of intense current research. For a 12- or 24-hour advance prediction of intensity changes, persistence forecasts are usually quite good.

However, all over the world today the final or official cyclone forecast, although aided by objective advisories like CLIPER or the output from a statistical-dynamical model, is a subjective decision where the skill and experience of the forecaster is most important. Fully deterministic prediction of cyclone intensity and movement, beginning from accurate initial data on the cyclone and its large-scale environment, using appropriate models and fast computers, is a goal being pursued in some centres of research, but its use in cyclone warning may be decades away in the future.

Errors (the distance between a forecast position and that subsequently occupied by the cyclone centre) in the forecasts prepared at the National Hurricane Center,

USA for the decade 1982–1991 averaged 100, 193 and 383 km respectively for 12-, 24- and 48-hour advance forecasts respectively (AMS 1993). All users of cyclone warnings should be aware of the current accuracy levels in forecasting cyclone tracks.

#### 4. Cyclone disaster mitigation

The cyclone distress mitigation committees constituted in the coastal states of India since 1969 have made steady progress in disaster mitigation efforts. A good example of its functioning may be seen from the action taken during a recent severe cyclone that crossed the Andhra coast on May 9, 1990. The feeble system that originated over the South-West Bay of Bengal intensified into a cyclonic storm by 1730 IST (Indian Standard Time) of May 5, 1990. It attained hurricane intensity on the morning of May 6 when its centre was 450 km South-East of Madras. Moving North-West it was located 300 km South-East of Madras on the morning of May 7. Moving further in a northerly direction it crossed the South Andhra coast near the mouth of the river Krishna (south of Machilipatnam) at about 1900 IST on May 9, 1990 and weakened.

The system was continuously tracked by INSAT-1B all through its life. It attained a peak intensity of 6.5 on the T-number scale on the morning of May 8. The system slightly weakened to T-5.5 before crossing the coast. After the cyclone came into radar range it was tracked by the cyclone detection radars at Karaikal, Madras and Machilipatnam during the period May 7 to 9.

The following information is extracted from a report on the cyclone prepared by the Andhra Pradesh State Government. On receipt of the first message regarding the cyclone by the government agencies (Chief Secretary, Collectors etc.) on May 5, a central Control Room (manned by representatives from revenue, police, army and telecommunication departments) was immediately activated. Subsequently, 28 more bulletins were received at the Control Room from the Meteorological Department. The Chief Minister of the State constantly reviewed the situation from May 6 onwards. A high-level meeting of cabinet ministers, senior secretaries to government and heads of departments was called by the Chief Minister on May 8 to review the action being taken. The cyclone control room was equipped with a police wireless system and HAM sets to maintain communication in the event of failure of normal communication links which generally get disrupted in a cyclonic storm situation. The assistance of voluntary agencies was sought. A high-level Natural Calamity Coordination meeting was also convened with the officers of the armed forces to formulate relief and rescue measures. In an exercise of an unprecedented nature, *6,51,865 persons from 546 villages were evacuated to safer places and sheltered in 1098 relief centres for 3 to 8 days*. The cyclone took a toll of 967 lives. But for the timely large-scale evacuation of people, the loss of life would have been colossal, as the storm surge associated with the cyclone, raised the sea level by four metres above the normal tide level (Subramanian and Mohanan 1992).

Cyclone distress mitigation has several components, but I would like to highlight (i) long-term measures, and (ii) measures to be adopted during a period of cyclone threat. Long-term measures include the construction of cyclone shelters to which humans and animals could be transported at short notice, the continuing education of officials and the public regarding cyclone hazards and the available warning and disaster mitigation systems, the periodic review of cyclone-warning terminology to

make the warnings easily and unambiguously understood by the public, and lastly coastal area planning to reduce property damage by cyclones. In all these areas except the last India has made notable advances.

According to expert opinion (Southern 1987, pp. 147–185) there are four ingredients necessary for implementing an effective system of long-term mitigation planning for coastal areas: (i) a technical evaluation of the climatological risk posed by cyclones and cyclonic effects in selected coastal sectors (Joseph *et al* 1981 is an attempt in that direction, prepared in connection with the planning stage for the establishment of a major steel mill near Paradeep); (ii) an assessment of the relative vulnerability of populations within these selected boundaries to the cyclonic effects of stated intensity or strength; (iii) The establishment of structural design codes, regulatory controls and minimum safety standards within an authoritative framework of legislation designed to encourage public adherence; and (iv) educational programmes to gain community acceptance of the costs of cyclone disaster mitigation.

Finally, in addition to the alerts, warnings and other advisory bulletins to the officials as per the CDMC plans, the public should be kept periodically informed about the progress of the cyclone, through all available media channels. The chief of the tropical cyclone warning centre should make a personal appearance on television, particularly at critical times during the approach of a cyclone so as to infuse confidence in the assessment made of the threat. Latest cyclone pictures taken by the radar network may be shown on such occasions to support the confidence in warnings. The public should be educated about the basic features, such as the appearance of a sudden lull in the wind as the eye of the storm passes over the station, to be followed by strong winds from an opposite direction about half an hour to an hour later. The public should also be advised not to heed rumours and to carefully listen to official bulletins over the radio and television and cooperate with the officials who act on directions from the Control Room which is in live contact with the related Area Cyclone Warning Centre.

## 5. Conclusions

India has now a fairly dependable and efficient cyclone warning system, comparable to those in USA and Australia.

The following suggestions for further improvement may be useful.

- Techniques for measurement of the low level wind field of the cyclone may be improved. Specifically,
  - (a) Robust wind instruments may be deployed along our coast that can withstand the fury of the cyclone and make accurate measurements up to at least 180 knots.
  - (b) The plan to replace the existing cyclone detection radars with Doppler radars may be expedited. In addition to monitoring the low level wind field, a Doppler radar can detect tornadoes in the field of cyclones. We have at present very little information on the occurrence of tornadoes in the Indian cyclones.
  - (c) Reconnaissance aircraft facility may be added to this system. It is suggested that we may begin with just two such aircraft (one in an operation mode and the other as a standby). Measurements made during these flights can greatly enhance the reliability of data obtained from satellites for inference of cyclone parameters including wind. Reconnaissance aircraft reports of cyclone location and intensity

are considered to be the most reliable. This facility can also be used for research on tropical cyclones.

- Quantitative information on rainfall rates can be derived from radar reflectivity mappings. When combined with digitally telemetered rain gauge reports, quantitative estimates of the rainfall in the cyclone field can be made. Radar provides a spatial and time resolution that cannot be achieved by a raingauge network. This is useful for assessing in real time, the potential flooding situations in association with cyclone rainfall (Elsberry 1987, pp. 1–12, 91–131). It would be prudent to develop this capability in India.
- Development of statistical–dynamical models linking cyclone motions with the observed large-scale atmospheric flow fields can help in forecasting cyclone motion up to 3 days in advance. But to accomplish this we have to develop capabilities for numerical prediction of the flow field for the same period. We have the potential to achieve this capability at the National Centre for Medium Range Weather Forecasting.
- Overwarning of our coastal areas in routine cyclone warning is common. The severe weather of a cyclone at any given instant affects a coastal area about 100 km in length. Since there are errors of the order of 200 km in a 24-hour forecast of the cyclone track, a severe weather warning normally covers 300–500 km of the coast, even in the case of cyclones with smooth tracks. To be effective warnings must be given 24 hours ahead so that there is adequate time for distress mitigation efforts particularly evacuations which can only be made during daylight hours. Considerably larger lengths of the coast-line can be included for warnings and alerts, particularly in cases of erratic movement of cyclones and cyclone motion along curved coastlines, since in the case of cyclones it is better to be overwarned than underwarned or taken unawares. The excellent capabilities to detect and track cyclones and to determine their intensities make such a warning a real assessment of the threat, unlike in the case of an earthquake which strikes without notice.
- We have seen that in the recent May 1990 cyclone about six and a half lakh persons were evacuated. It was a necessary exercise under the then existing conditions to reduce loss of life, but shifting people large distances away from their homes brings untold misery to them, and it is realised that lateral evacuation to distant sites may not be the most appropriate solution. WMO has recommended development of *in-situ* survival strategies: Construction of adequate cyclone survival shelters suitable for vertical evacuation. With such an option, warning times could be reduced, thereby minimising overwarning (WMO 1990).
- Making available robust communication facilities at the time of cyclones is an important requirement for India. At present we have an operational Disaster Warning System (DWS) linked with INSAT which was developed and implemented in the coastal districts of Tamilnadu and Andhra Pradesh in 1987 for rapid and direct dissemination of cyclone warnings from the Area Cyclone Warning Centre at Madras to a large number of receiving stations. Raghavan & Sen Sarma (1989) report that the system was adjudged to be quite successful on all the three occasions when cyclones struck our coasts in 1987. There are plans to extend this facility to other states and this may be expedited.

I thank the Indian Academy of Sciences for inviting me to contribute to the special issue of *Sadhana*. Thanks are also due to Dr N Sen Roy, New Delhi, Dr G S Mandal, New Delhi, Sri S Ganesan, Madras, Sri S K Subramanian, Madras, and Sri S Raghavan, Madras for helpful discussions. The radar picture of the cyclone in figure 1 was kindly provided to the author by the India Meteorological Department, Government of India.

## References

- AMS 1993 Policy statement on hurricane detection, tracking and forecasting of the American Meteorological Society, as adopted by the Executive Committee on 23 April 1993. *Bull. Am. Meteorol. Soc.* 74: 1377-1380
- Anthes R A 1982 Tropical cyclones, their evolution, structure and effects. *Meteorological monographs* (Boston: Am. Meteorol. Soc.) vol. 19, no. 41
- Dvorak V F 1975 Tropical cyclone analysis and forecasting from satellite imagery. *Mon. Weather Rev.* 103: 420-430
- Elsberry R L (ed.) 1987a Observation and analysis of tropical cyclones. *A global view of tropical cyclones* (Monterey, CA: Naval Postgraduate School)
- Elsberry R L (ed.) 1987b Tropical cyclone motion. *A global view of tropical cyclones* (Monterey, CA: Naval Postgraduate School)
- Gray W M 1979 Hurricanes, their formation, structure and likely role in tropical circulation. *Meteorology over the tropical oceans* (ed.) D B Shaw (Bracknell: R. Meteorol. Soc.)
- Gray W M, Neumann C, Tsui T J 1991 Assessment of the role of aircraft reconnaissance on tropical cyclone analysis and forecasting. *Bull. Am. Meteorol. Soc.* 72: 1867-1883
- Joseph P V, Ghosh S K, Sharma B L, Seshadri V R 1981 Report on the impact of cyclonic storm and storm surge near Paradeep. India Meteorological Department
- Landsea C W 1993 A climatology of intense (or major) Atlantic hurricanes. *Mon. Weather Rev.* 121: 1703-1713
- Mishra D K, Gupta G R 1976 Estimation of maximum wind speeds in tropical cyclones occurring in Indian seas. *Indian J. Meteorol. Hydrol. Geophys.* 27: 285-290
- Raghavan S, Sen Sarma A K 1989 DWS- A satellite based warning system. *Proceedings of the symposium on Preparedness, Mitigation and Management of Natural Disasters*, pp. 4-27 to 4-30.
- Southern R L 1979 The global socio-economic impact of tropical cyclones. *Aust. Meteorol. Mag.* 27: 175-195
- Southern R L 1987 Tropical cyclone warning and mitigation systems. *A global view of tropical cyclones* (ed.) R L Elsberry (Monterey, CA: Naval Postgraduate School)
- Subramanian S K, Mohanan G 1992 Some features of the Bay cyclone of May 1990. *Vayu Mandal* 22: 97-102
- WMO 1990 Tropical cyclone warning systems WMO/TD No. 394, Report No. TCP-26, World Meteorological Organisation, Geneva

# Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories

U C MOHANTY

Centre for Atmospheric Sciences, Indian Institute of Technology, Hauz Khas,  
New Delhi 110 016, India

Present address: National Centre for Medium Range Weather Forecasting,  
IMD Complex, Lodi Road, New Delhi 110 003, India

**Abstract.** Based on information about tropical storms and depressions in the Bay of Bengal over a 100-year period (1877–1976), certain climatic characteristics of tropical cyclones are examined. A brief description of climatic parameters, notably the region of their development, direction and nature of movement and percentage of disturbances intensified into severe tropical storms in different seasons of the year, which are explicitly or implicitly related to the formulation of objective methods for forecasting storm tracks, is presented in this work.

A brief review of objective methods (numerical and statistical) used for forecasting the tracks of cyclonic storms and the rationale for development of such methods are described. Two numerical methods based on integral and steering flow concepts and four different physical-statistical methods, which take into consideration the influence of both external and internal forces responsible for the movement of tropical cyclones, are developed for forecasting the trajectories of post monsoon cyclonic storms in the Bay of Bengal.

Performance of these six objective methods are illustrated through the study of a homogeneous sample of cyclonic storms (14 cases) in the Bay of Bengal during the post monsoon seasons of 1975 and 1976. These results are compared with those obtained from some of the existing methods for forecasting the movement of tropical cyclones in the Bay of Bengal. Finally, the limitations and prospects of objective methods in predicting the tracks of cyclonic storms in the Bay are discussed.

**Keywords.** Bay of Bengal; tropical cyclones; storm tracks; deterministic methods; numerical weather prediction.

## 1. Introduction

The Bay of Bengal is a potentially energetic region for the development of cyclonic storms, accounting for about 7% of the global annual total number of tropical storms (Gray 1968). Furthermore, the Bay storms are exceptionally devastating, especially



when they cross over into the coastal states of India and Bangladesh (De Angelis 1976). However, inadequate present-day knowledge about their physical mechanisms of development, evolution and movement and deficient observational data close to the radius of influence of cyclonic storms limit the degree of physical and mathematical sophistication and thus make it difficult to develop more accurate objective methods for forecasting their movements.

Occurrence of cyclonic storms in the Bay of Bengal continues to be of great concern to India and Bangladesh and prediction of their tracks is one of the most challenging problems in meteorology.

The purpose of the present study is to investigate the problem of predicting the tracks of tropical storms in the Bay of Bengal using objective (numerical and statistical) techniques. First, a brief account of certain climatic characteristics of tropical cyclones in the Bay of Bengal are presented which provide some guidelines for the formulation of objective methods. Compared to other seasons, the post-monsoon season storms in the Bay of Bengal are more frequent and devastating in nature and a fair number of them have recurving tendencies, which make the forecasting of their evolving trajectories extremely difficult. Two numerical and four statistical methods have been developed for forecasting the track of post-monsoon season tropical storms. Performances of these six objective methods are illustrated, based on a homogeneous sample of cyclonic storms that occurred in the Bay of Bengal during the years 1975 and 1976.

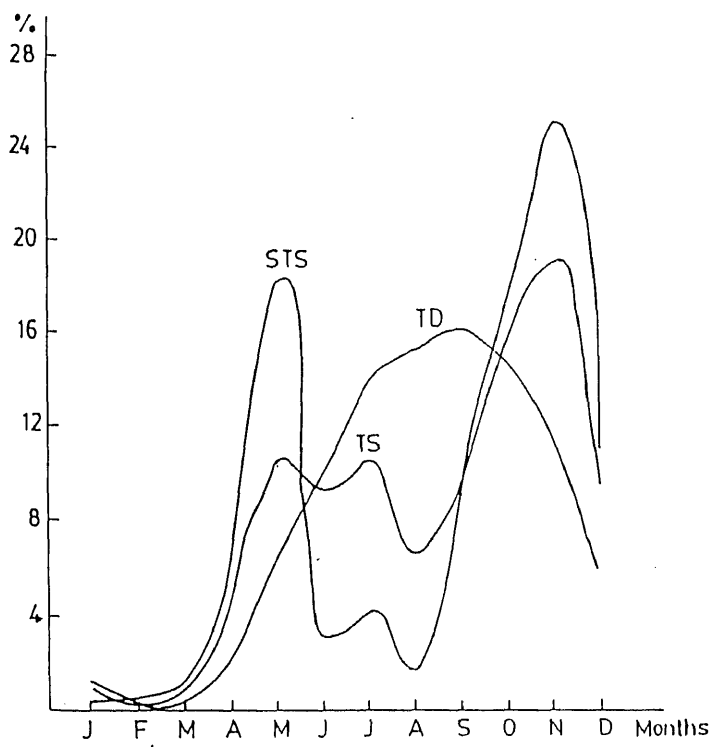
## 2. Climatology of tropical cyclones

Climatology of the development and movement of cyclonic storms is quite helpful in the process of formulating objective methods for forecasting their tracks. The frequency, location of the initially observed storms and nature of their movements are some of the important factors which are considered in this work.

For this study, information about the tracks of the storms and depressions in the Bay of Bengal for 100 years (1877–1976) were collected from India Meteorological Department publications. Data for the first 94 years were obtained from the India Met. Dept. (1979) atlas and for the remaining 6 years (1971–1976) from other publications.

Cyclonic disturbances are classified into three categories as Tropical Depressions (TD), Tropical Storms (TS) and Severe Tropical Storms (STS) depending on their maximum sustained winds by various systems as defined by the India Meteorological Department (India Met. Dept. 1979). In this study TD stands for all the cyclonic disturbances which attained the stage of a tropical depression (TD) and thus includes TS and STS. TS includes all cyclonic disturbances in which the maximum sustained winds were 34 kt (i.e. 63 km/h) or higher. Further, a year was divided into four seasons, each of 3 months duration. The period January to March is a passive period for cyclonic disturbances. April to June is considered the pre-monsoon period, July to September, the monsoon, and October to December, the post-monsoon period, though the actual monsoon season lasts from June to September.

Figure 1 shows the annual percentage distribution of TD, TS and STS for a period of 100 years. More than 60% of all cyclonic disturbances occur between July and October with a maximum in September, whereas only about 2% occur between January and March. The cyclonic storms show three maxima, with the primary one



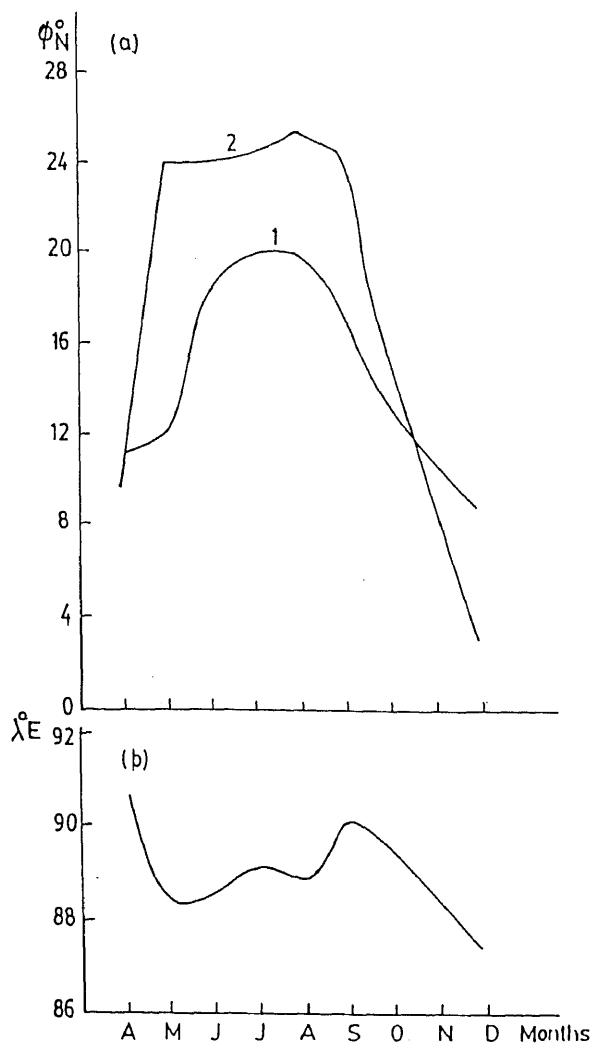
**Figure 1.** Percentage distribution of the occurrence of tropical disturbances (TD), tropical storms (TS) and severe tropical storms (STS) relative to calendar year.

occurring in the month of November (19.5%). Severe cyclonic storms exhibit two sharp maxima with the primary one (more than 25%) occurring in November and another (about 19%) in May.

Some of the interesting conclusions based on these results are as follows :

- i) January to March is a passive period for cyclonic disturbances and thus will not figure in further discussions.
- ii) The most active period of tropical depressions (TD) is the summer monsoon period (July–September). However, this period is quite passive for the occurrence of TS and STS, as only 2 to 7% of the depressions intensify into TS and STS.
- iii) The maximum number of cyclonic storms (about 46%) occur in the post-monsoon period (October–December). This is also the period of generation of the maximum number of severe tropical storms (more than 54%) in the Bay of Bengal.
- iv) In general, there are two severe cyclonic storm seasons in the Bay of Bengal. The primary one is the post-monsoon period (54.5%) and the second, the pre-monsoon season with 27.6% STS.

Figure 2 illustrates seasonal variations of the location (latitude- $\phi$  and longitude- $\lambda$ ) at which initial cyclonic disturbances, which later became TS and STS, were detected. It is seen that the location of initially observed cyclonic disturbances is further northward ( $18^{\circ}$ – $20^{\circ}$ N) during the monsoon season compared to the pre- and post-monsoon periods ( $8^{\circ}$ – $12^{\circ}$ N). Such a latitudinal migration of the location of the occurrence of cyclonic storms agrees well with the latitudinal migration of the equatorial trough



**Figure 2.** Seasonal (a, 1) latitude and (b) longitude variations of initially located cyclonic disturbances which later became tropical storms. (a, 2 corresponds to the seasonal latitude variation of the equatorial trough from Gray 1968.)

(Gray 1968) and the inter-tropical convergence zone (Minina & Niguen 1969). Seasonal longitudinal variations of the location of cyclonic disturbances which later became TS and STS is not large ( $87\text{--}90^\circ E$ ). From figures 1 and 2 (curve-1), we arrive at the interesting conclusion that if a cyclonic disturbance originates in the Central Bay of Bengal ( $9\text{--}10^\circ N$  and  $88\text{--}90^\circ E$ ), then the probability of its intensification into TS and STS becomes very high (40 to 80%). On the other hand, the probability of intensifications of a cyclonic disturbance into TS and STS decreases considerably (2 to 15%) if it initially forms in the northern part of the Bay ( $18\text{--}20^\circ N$  and  $88\text{--}89^\circ E$ ).

For further detailed study, computation of the mean ( $\mu$ ), standard deviation ( $\sigma$ ), coefficient of variation ( $cv$ ), skewness ( $S$ ), kurtosis ( $K$ ) and also standard error of skewness ( $\sigma_s$ ) and kurtosis ( $\sigma_k$ ) of the initial location (latitude and longitude) of cyclonic disturbances which intensified into TS and STS, were carried out by Mohanty & Dube (1981) with a 100-year data set. The results are given in table 1.

**Table 1.** Statistical characteristics of the initial location of the cyclonic disturbances, intensified into tropical storms and severe tropical storms (1877-1976).

Period	Latitude ( $\phi^\circ$ N)					Longitude ( $\lambda^\circ$ E)						
	$\mu$	$\sigma$	$cv$	$S$	$K$	$\mu$	$\sigma$	$cv$	$S$	$K$	$\sigma_s$	$\sigma_k$
Pre-monsoon (IV-VI)	15.0	4.06	0.27	-0.06	-1.19	88.9	3.50	0.04	0.01	-0.28	0.57	0.45
Monsoon (VII-IX)	18.8	2.51	0.13	-1.17	2.05	89.4	2.32	0.03	1.02	3.55	0.54	0.43
Post-monsoon (X-XII)	11.2	3.19	0.29	0.39	-0.27	88.6	3.99	0.05	0.35	-0.32	0.42	0.34
April-December (IV-XII)	14.2	4.56	0.32	0.02	-1.18	88.9	3.49	0.04	0.23	0.02	0.29	0.24

**Table 2.** Percentage frequency distribution of the nature of the trajectories of cyclonic storms (1877-1976).

Type of traj.	Period				Type of traj.	Period			
	Pre-mons (IV-VI)	Mons (VII-IX)	Post-mons (X-XII)	Apr-Dec (IV-XII)		Pre-mons (IV-VI)	Mons (VII-IX)	Post-mons (X-XII)	Apr-Dec (IV-XII)
W	3.6	9.0	7.8	7.1	NE	13.5	0.8	8.8	7.8
WNW	6.3	24.0	11.2	13.5	ENE	0.9	—	—	0.2
NW	9.9	30.6	16.1	18.5	E	0.9	—	—	0.2
NNW	18.1	17.4	7.3	12.8	SCT	16.2	7.4	34.6	22.4
N	15.3	6.6	5.4	8.3	MCT	0.9	2.5	2.9	2.3
NNE	14.4	1.7	5.9	6.9					

Abbreviation: Traj. - trajectories; mons. - monsoon; Apr. - April; Dec. - December.

From the results in table 1, the following conclusions are drawn :

- (i) During the monsoon period, latitudes of initially observed cyclonic disturbances, which later became TS and STS, are more homogeneous compared to the pre- and post-monsoon periods, as both the standard deviation ( $\sigma$ ) and the coefficient of variation ( $cv$ ) of  $\phi$  during the monsoon period are the least ( $2.51^\circ$  and  $0.13\%$  respectively).
- (ii) The coefficient of variation of longitude is very small (less than  $0.05\%$ ) during all the three seasons. This implies that the longitude of the origin of cyclonic disturbances is more homogeneous compared to the latitude.
- (iii) From skewness and kurtosis of  $\phi$  and  $\lambda$  and their standard errors, it is clear that during the post-monsoon period, the skewness and kurtosis of both  $\phi$  and  $\lambda$  do not exceed  $\sigma_s$  and  $\sigma_k$  respectively. Initial locations of the post-monsoon cyclonic disturbances which later intensify into TS and STS are thus symmetrical about the mean position ( $11.2^\circ\text{N}$  and  $88.6^\circ\text{E}$ ) and follow the normal laws of distribution.

According to the nature of their trajectories, TS and STS may be broadly divided into two categories, namely, the storms with simple trajectories (if movement of TS and STS is almost in one direction) and those with parabolic trajectories (i.e. movement of TS and STS in a recurved path). About 75% of the cyclonic storms in the months of April to December belong to the first category and the remaining 25% storms exhibit tracks with one or more points of recurvature.

For a detailed account of the nature of movement of cyclonic storms, those belonging to the first category were further divided into 9 classes according to their direction of movement on 16 points of the compass as they do not move into the southern sector (i.e. West-South West to East-South East). Cyclonic storms of the second group were divided into two classes: storm tracks with a single point of recurvature (SCT) and those with two or more points of recurvature (MCT).

Table 2 gives the seasonal percentage frequency of the nature of trajectories of cyclonic storms which are divided into 11 classes as stated above. The results in table 2 lead to the following conclusions:

- (i) During April to December, the maximum number of cyclonic storms (22.4%) exhibit tracks with a point of recurvature and a smaller number (18.5%) move in the North-West (NW) direction. Over a 100-year period, the number of cyclonic storms which moved in the East-North East (ENE) and East (E) directions were only two (about 0.4%) whereas those that moved in the West-North-West (WNW), North-West (NW) and North-North-West (NNW) directions constituted more than 45%.
- (ii) The nature of the tracks of TS and STS vary from season to season. In the pre-monsoon period, more than 60% of the storms move in the NWN, N, NNE and NE directions, with a maximum number of storms (18.1%) in the NNW direction. During the monsoon period, more than 80% of the cyclonic storms move in W, WNW, NW and NNW directions, the majority (30.6%) in the NW direction.
- (iii) Maximum number (34.5%) of cyclonic storms occur in the post-monsoon period and over 72% of these are observed to recurve their tracks. Only 18% and 10% of such storms occur in the pre-monsoon and monsoon periods respectively.

These results on the climatology of tropical cyclones in the Bay of Bengal show

that the maximum number of tropical storms (more than 46%) and severe tropical storms (more than 54%) occur in the post-monsoon period. Also the maximum number of (more than 72%) recurved tracks of cyclonic storms appear during this period of the year.

The task of predicting the recurvature of a storm's track constitutes a difficult problem and most of the forecasting models fail at this point. Therefore an attempt has been made in this work to develop some objective techniques to predict the movement of cyclonic storms in the Bay of Bengal during the post-monsoon period.

### **3. Objective methods of forecasting the movement of cyclonic storms**

Objective methods have made significant contributions to the theory and practice of weather forecasting on all time scales. In general, objective methods are divided into two classes : numerical (i.e. dynamical) methods and statistical methods. Both these methods are widely used to provide routine objective guidance on prediction of various atmospheric phenomena. Though dynamical methods are based on the laws of physics and mathematical theories and are found to be quite successful in forecasting large-scale atmospheric processes in the mid-latitudes, their performance is not very satisfactory in predicting the movement of tropical cyclones. This is mainly due to inadequate information available about tropical cyclones as well as lack of adequate understanding of the physical mechanism of their development, evolution and movement. Performance of present day statistical methods in predicting the movement of tropical storms is therefore better or at least comparable with those of numerical methods. Further, statistical methods are simple and economical. As such, a survey of the worldwide objective prediction models by Hope & Neumann (1977) cites 31 operational models, most of which are statistical models. However, future progress in prediction models to provide objective guidance for routine forecasts of the tracks of cyclones will depend heavily on dynamical methods. Clearly the most reliable forecasts in the future will be produced by approaches or procedures that make judicious use of both numerical and statistical methods.

A brief account of a number of dynamical and statistical methods, developed to predict the tracks of post-monsoon cyclonic storms in the Bay of Bengal, is given below.

#### **3.1 Numerical methods**

Based on the concept of the circulation patterns in and around a tropical cyclone, the numerical (i.e. dynamical) methods for forecasting the movement of storm tracks are generally divided into two broad groups, the steering flow method (external forces) and the integral method (internal as well as external forces). In this work attempts have been made to present two dynamical models which belong to the above two groups.

Further, an attempt has also been made to present more complex numerical models based on multi-level primitive equations with appropriate physical processes incorporated in parameterization form. Such models are used for the prediction of synoptic scale atmospheric processes including intense atmospheric vortices (such as cyclones, depressions, western disturbances etc). However, such models require a knowledge of more comprehensive, dense and high quality meteorological fields as initial conditions for integration.

**3.1a Steering flow method:** In this method, the cyclonic storm is represented by a constant circular vortex stream described in terms of the maximum sustained wind and the radius of the storm's eye of its influence. Also barotropic flow is assumed. Further, the vortex of the tropical cyclone is eliminated from the initial stream function analysis leaving only a steering wind field in the region of the cyclone. The cyclonic vortex motion is estimated by the steering current which is determined by a numerical model. This concept was used for forecasting the tracks of cyclonic storms by Kasahara (1959), Sitnikov (1968), and others.

A brief description of this method is given here and one may refer to Kivganov & Mohanty (1979b) for details. In this method the total barotropic stream function ( $\psi$ ) is considered to be the sum of a steering stream function ( $\bar{\psi}$ ) and a local cyclonic vortex stream function ( $\psi^*$ ), that is

$$\psi = \bar{\psi} + \psi^*. \quad (1)$$

Further,  $\psi^*$  is divided into two components as

$$\psi^* = \psi_1^* + \psi_2^*, \quad (2)$$

where  $\psi_1^*$  is the vortex stream function due to tangential wind  $V_0$  and  $\psi_2^*$  is the stream function due to translational motion  $v_0$  of the storm. For a detailed description of the computation of  $\psi_1^*$  and  $\psi_2^*$ , one may refer to Kivganov & Mohanty (1979b).

The total barotropic stream function ( $\psi$ ) is computed from the horizontal components of the wind (Mohanty & Madan 1983), and the steering stream function ( $\bar{\psi}$ ) is estimated by (1).

The instantaneous speed of the movement of the centre of a cyclonic storm can be computed from the steering stream function ( $\bar{\psi}$ ) by the following relations (Mohanty 1978).

$$\begin{aligned} C_{x_0} &= -m \frac{\partial \bar{\psi}}{\partial y} - \frac{m}{\mu^2} \frac{\partial}{\partial y} (m^2 \nabla^2 \bar{\psi} + f), \\ C_{y_0} &= m \frac{\partial \bar{\psi}}{\partial y} - \frac{m}{\mu^2} \frac{\partial}{\partial y} (m^2 \nabla^2 \bar{\psi}), \\ V_0 &= (C_{x_0}^2 + C_{y_0}^2)^{1/2}, \end{aligned} \quad (3)$$

where  $C_{x_0}$  and  $C_{y_0}$  are the speed of movement of the centre of the storm along the  $x$  and  $y$ -axes respectively,  $m$  is the map factor,  $f$  is the Coriolis parameter, and  $\mu$  is a characteristic parameter of the horizontal dimension of the storm ( $\mu R = 3.83$ ) and  $\nabla^2$  is the Laplacian operator.

For forecasting the movement of the cyclonic storm using (3), the steering stream function is predicted by a modified non-divergent barotropic model,

$$(m^2 \nabla^2 - \alpha^2)(\partial \bar{\psi} / \partial t) = m^2 J(m^2 \nabla^2 \psi + f, \bar{\psi}) + 9m^2 \nabla^2 (m^2 \nabla^2 \bar{\psi}), \quad (4)$$

where  $\alpha^2$  is a constant ( $\alpha^2 = f^2 / R \bar{T}$ ),  $R$  the gas constant for air,  $\bar{T}$  the mean tropospheric temperature,  $9$  the kinematic coefficient of eddy viscosity and  $J$  the Jacobian operator.

This method for forecasting the movement of the centre of the cyclone involves two steps. At each time step a predicted value of the steering stream function ( $\bar{\psi}$ ) is obtained by solving the modified vorticity equation (4). In the next step the predicted

value  $\bar{\psi}$  is used to obtain the instantaneous speed and direction of the movement of the centre of the storm by relation (3).

**3.1b Integral method:** In this method, the cyclonic storm is considered an inseparable part of the large-scale flow and thus free interaction of the cyclonic storm with its outer environment is allowed. The predicted path of the tropical storm results from tracking its centre as a minimum geopotential height/stream function value at a constant pressure level. This method is widely used for forecasting the movement of tropical storms (Sanders *et al* 1975, Sikka 1975, Singh & Saha 1976, and others). In general either barotropic vorticity equation or primitive equation forecasting model is used for the integral method.

In the present study we have used the primitive equation barotropic model for forecasting the movement of post-monsoon storms in the Bay of Bengal. The model equations, initialization procedure and numerical solution of this model are described in an earlier work by Mohanty (1982). This model is used to predict the geopotential height and horizontal wind fields. Based on the analyses of these forecasted fields the centre of the storm and thus its future trajectory are determined.

**3.1c Complex atmospheric models:** In recent years, with the advances in high speed computers, more complex atmospheric models have been developed for simulation/prediction of large-scale atmospheric phenomena. They are broadly of two categories: (i) Limited area models (LAM) for a specified region with a capability to integrate over a shorter period of 1–2 days to obtain short-range prediction of the regional features in greater detail, and (ii) general circulation models (GCM) for the entire globe with capability to integrate for any length of time as they are not influenced by artificial lateral boundary conditions. GCM is generally used for extended range simulation/prediction over the entire globe.

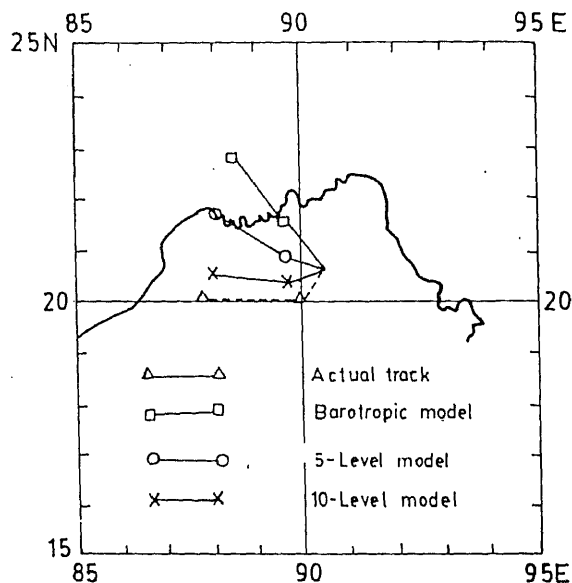
Both LAM and GCM can be used to predict the movement of tropical cyclones and other atmospheric vortices (such as lows, depressions and deep depressions) based on the concept of integral methods (as discussed in § 3.1b).

For the sake of completeness a brief description of each of the LAM and GCM is presented. As an example of the performance of these models, an illustration of the prediction of storm tracks is also presented.

(i) *Limited area numerical weather prediction model* (Mohanty *et al* 1989) – In this study a primitive equation (PE) model governing the motion of the atmosphere, appropriate to a meso-scale quasi-hydrostatic baroclinic system is considered. The model with five prognostic and two diagnostic equations (viz. the zonal and meridional momentum, the thermodynamic, the moisture continuity, the surface pressure tendency, the hydrostatic and the mass continuity equations) for the seven basic meteorological variables forms a closed system and the model equations are represented in flux form, in spherical coordinates. The model uses sigma coordinates in the vertical and incorporates topography and physical parameterizations of the planetary boundary layer, air–sea exchange of heat and moisture and precipitation processes. Details of the model as well as numerical solution of the model equations and initialization procedure of mass and velocity fields in the model are presented in earlier work by Mohanty *et al* (1988, 1989).

As an illustration of the performance of the model, the actual track of a monsoon depression in the Bay of Bengal during intensive observation periods of the monsoon





**Figure 3.** Performance of the barotropic and multi-level (5-layer and 10-layer) primitive equation regional models (LAM) to predict the track of a monsoon depression over the Bay of Bengal.

experiment (MONEX)-79 from July 1979, and the corresponding forecast tracks of the disturbance obtained from the barotropic PE model (described in §3.1b) and the multi-level PE model with 5 and 10 vertical levels, are shown in figure 3. In all these cases, the horizontal domain of the model is bounded by  $30^{\circ}\text{E}$ – $105^{\circ}\text{E}$  and  $7.5^{\circ}\text{S}$ – $45^{\circ}\text{N}$  with a grid resolution of  $1.875^{\circ}$ . For this study the FGGE level-IIIb analysis of the European Centre for Medium Range Weather Forecasts, UK obtained as a result of the First GARP Global Experiment (FGGE) and MONEX-79, which is considered to be the first-ever excellent database for numerical studies on the summer monsoon, was obtained to constitute the initial conditions. Track prediction based on the integral method with barotropic, 5-level and 10-level PE models clearly demonstrates the impact of a better representation of realistic atmosphere and surface processes including topography on the improvement of such forecasts.

(ii) *General circulation model (GCM)* – General circulation models over the entire globe do not require lateral boundary conditions. Such models are generally used for extended range simulation/prediction of the atmospheric processes. In this study, the operational global model of the European Centre for Medium Range Weather Forecasts (ECMWF), UK was used to illustrate its performance in the prediction of storm tracks in the Bay of Bengal. A detailed description of the T-63 spectral GCM is given by Simmons & Jarraud (1984) and performance of the model with modified parameterization of the physical processes in predicting the 1979 summer monsoon onset is presented by Slingo *et al* (1988). The model is used to predict the track of two monsoon depressions in the Bay of Bengal during 23 June–26 June, 1979 and 28 June–1 July, 1979. The observed and predicted tracks are shown in figure 4. For this study, the FGGE level-IIIb data set was used as initial conditions. While the forecasts for the tracks of monsoon disturbances upto 72 hours are reasonably close to the observed tracks, a large difference in the determination of initial positions of the disturbances is mainly attributed to lack of adequate and accurate observations in their vicinity in data-sparse oceanic regions.

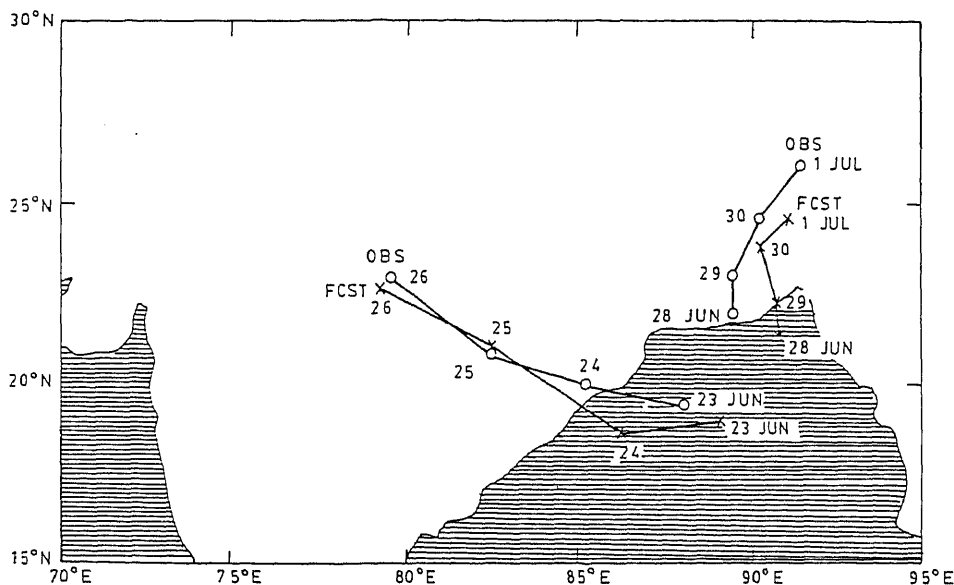


Figure 4. Performance of a multi-level global spectral model (GCM) to predict the track of a monsoon depression over the Bay of Bengal.

3.1d *Limitations and prospects of numerical methods:* Considerable success has been achieved in predicting large scale atmospheric motions using numerical weather prediction models, largely owing to the availability of high-speed supercomputers, space-based observation systems, data assimilation, initialization, sophisticated numerical techniques and better understanding of various physical processes and their parameterization in Numerical Weather Prediction (NWP) models. However, improvement in prediction of tropical cyclones using NWP models has been sluggish mainly due to lack of accurate and adequate data coverage over tropical oceans, which are the breeding grounds of tropical cyclones. Furthermore, we do not as yet have a clear understanding of their thermodynamics, particularly, parameterization of convection, cloud-radiation, boundary layer, air-sea interaction and their feedback processes. Future improvements in NWP for prediction of the tracks of tropical cyclones/disturbances actually call for the following developments.

- (i) Augmented coverage of observing systems over tropical oceans.
- (ii) Optimum utilization of non-conventional data sets on tropical cyclones obtained from satellites, weather radar and aircraft etc. in the preparation of initial data for NWP models.
- (iii) Production of synoptic data sets on wind, temperature and moisture in the vicinity of tropical cyclones from their observed characteristic features (synthetic data generation).
- (iv) Appropriate diabatic and physical initialization procedures to retain/strengthen the realistic intensity of tropical cyclones (better representation of divergent flow and moisture fields).
- (v) Nudging the centre of the observed cyclones with that of the objectively analysed centres to be served as initial conditions for NWP models.

- (vi) Use of realistic sea surface temperature and topography data as lower boundary conditions in the NWP models.
- (vii) Development of high resolution regional/global models with horizontal resolutions of about 30–40 km and discretized for 25–30 vertical levels, for better representation of tropical cyclones in the NWP models.
- (viii) Improvement in the parameterization of the physical processes in the NWP model, in particular, convection, cloud radiation, boundary layer and air–sea exchange processes and their feedback mechanisms in association with tropical cyclones.

### 3.2 Statistical methods

Tropical storms move under the influence of both external and internal forces. External forces refer to the effect of large-scale atmospheric processes on their movement which is mainly manifested through steering flow. Internal forces arise from the storm itself and from its interaction with the steering flow in the troposphere.

Different techniques and predictors are used to consider these forces directly or indirectly in formulating statistical methods for forecasting the movement of tropical storms. In general, linear and nonlinear regression analyses are widely used. Potential predictors may be classified into four categories: synoptic, inertial, climatological and those obtained from large-scale meteorological fields through empirical functions (like empirical orthogonal function). Statistical methods differ from one another mainly by the statistical techniques and predictors used in their formulation.

In the present work, we propose four different statistical methods to predict the movement of post-monsoon season storm tracks in the Bay of Bengal. A brief description of these four methods is given below.

*Method 1 (Mohanty 1979):* This method is formulated on the basis of synoptic predictors. The sea level pressure field ( $P_0$ ), geopotential fields at 700 and 500 hPa pressure levels ( $H_{700}$  and  $H_{500}$ ) and the 1000 to 700 hPa and 700 to 500 hPa thickness are the current synoptic fields which are obtained over a regular  $5^\circ$  latitude/longitude moving coordinate,  $7 \times 8$  grid system, at the centre of a cyclonic storm. Other parameters taken into consideration are geostrophic steering currents at 700 and 500 hPa surfaces and the preceding 12 hour track of the storm centre ( $S_x$ ,  $S_y$ ). A stepwise screening procedure is used to obtain a series of regression equations for advance computations of the zonal and meridional motions for periods of 0–12, 12–24, 24–36 and 36–48 h.

*Method 2 (Kivganov & Mohanty 1978):* This method makes use of the first five coefficients of empirical orthogonal functions (e.o.f.) of the meteorological fields  $P_0$ ,  $H_{700}$ ,  $H_{500}$  and 100 to 700 and 700 to 500 hPa thickness as potential predictors, instead of the actual meteorological fields. Additional prediction parameters used are the empirical orthogonal functions of the e.o.f. coefficients of  $P_0$ ,  $H_{700}$  and  $H_{500}$ . Such a second expansion of e.o.f. allows one to incorporate the interaction in vertical co-ordinates. Further, the persistence of the systems  $S_x$  and  $S_y$  are also used as predictors. A set of prediction equations similar to those of method 1 are obtained by stepwise regression analysis.

*Method 3 (Kivganov & Mohanty 1979a):* This method is based on empirical predictors only (i.e. climatology and persistence) and is similar to those developed for other

regions (Neumann 1972, Hope & Neumann 1973, Neumann & Randrianarison 1976, and others). Seven basic predictors used in this scheme are: day number, latitude and longitude of the centre of the storm, corresponding to current time  $t_0$ , and average zonal and meridional speed of the cyclone over the previous 12 and 24-hour period. Third-order polynomial equations, consisting of seven primary and 112 secondary predictors are subjected to stepwise screening procedures to obtain a set of prediction equations as in methods 1 and 2.

*Method 4 (Mohanty 1980):* This method is based on the principles discussed in the work of Neumann *et al* (1972). First, three independent sets of forecasts, derived from the above three methods (methods 1–3) are obtained. These are then used as basic predictors to derive a new set of prediction equations. In order to consider the nonlinear effects, a third-order polynomial with 3 basic predictors and 16 additional predictors are generated. A stepwise screening regression procedure is used to eliminate the predictors which fail to provide the prescribed minimum incremental reductions of variance (0.1%). Prediction equations are derived from different sources as used in methods 1, 2 and 3.

#### 4. Data

For this work, using dynamical models we used vertically averaged wind fields at 850, 700, 500, 300 and 200 hPa pressure levels. The geopotential field was derived from wind fields by calculating stream functions and subsequently by reverse balance equations. Such a procedure for obtaining the geopotential field is found to be quite satisfactory (Mohanty & Madan 1983). Data were obtained over a rectangular region ( $3^{\circ}$ – $40^{\circ}$ N and  $60^{\circ}$ – $120^{\circ}$ E) with a regular grid of length 210 km for 14 cases of post-monsoon storms in the years 1975 and 1976. Five-year data samples of cyclonic storms from 1970 to 1974 were used to obtain prediction equations using statistical methods 1 and 2, while 98-year climatological data (1877–1974) were used for method 3. For method 4, the predicted values of zonal and meridional motions obtained by statistical methods 1, 2 and 3, above, were used on a homogeneous, reliable data set, consisting of 40 cases of post-monsoon storms (1970 to 1974). Performance tests of these four statistical methods were carried out on an independent data sample, comprising 14 cases of post-monsoon TS and STS in 1975 and 1976. In this study, a case of the storm is referred to 2-day trajectories. Thus, if a cyclonic storm continues for a period of 3 days, it provides 2 cases for study of 48-hour forecasts (day 1–2 and day 2–3). The data requirement and data set used for multi-level NWP models are discussed in §3.1 c.

#### 5. Results and discussion

Performance of the objective methods is measured in terms of the distance between the forecast and observed positions of the centre of the storm, referred to as the error vector ( $\Delta\mathbf{R}$ ). The mean vector error is estimated by:

$$\Delta\mathbf{R} = \frac{1}{N} \sum_{i=1}^N (\Delta CX_i^2 + \Delta CY_i^2)^{1/2}, \quad (5)$$

where  $N$  is the number of cases and  $\Delta CX$  and  $\Delta CY$  are zonal and meridional errors respectively.

Here, our study will be confined to only two dynamical models: the steering flow model (§3.1a) and the barotropic integral model (§3.1a) to compare their performances with those of the four statistical methods (§3.2). As adequate data sets were not available for more complex dynamical modelling (§3.1) for 1975–76, these models will not be considered here. The performance of the complex LAM and GCM have already been described in §3.1c in respect of tropical disturbances with FGGE level III-b data sets of 1979.

In order to investigate the performance of the six objective methods discussed in §3, mean vector errors were obtained from 14 cases of the post-monsoon storms in the Bay of Bengal over a two-year period (1975–1976) for different forecast time intervals as illustrated in figure 5. Clearly, the performance of the statistical method 4 is better as compared with those of others over all the time intervals (0–12 to 36–48 hours) of forecast. Performance of the integral method is quite satisfactory and better than those of the remaining four methods examined in this study. It is found that in relative terms, performance of the statistical methods in shorter time intervals of up to 24 hours is superior to those of numerical methods, but for longer time intervals the dynamical methods prove better.

Performances of the six objective methods, proposed in this work, for forecasting the movement of the post-monsoon cyclones in the Bay of Bengal were compared with some of the existing models for this region (Sikka & Suryanarayan 1972; Bansal & Dutta 1974; Sikka 1975) and found to be quite satisfactory. However, such a comparison is not truly representative as these other methods treated both pre- and post-monsoon storms and used different sample lengths. Prediction of the track with a point of

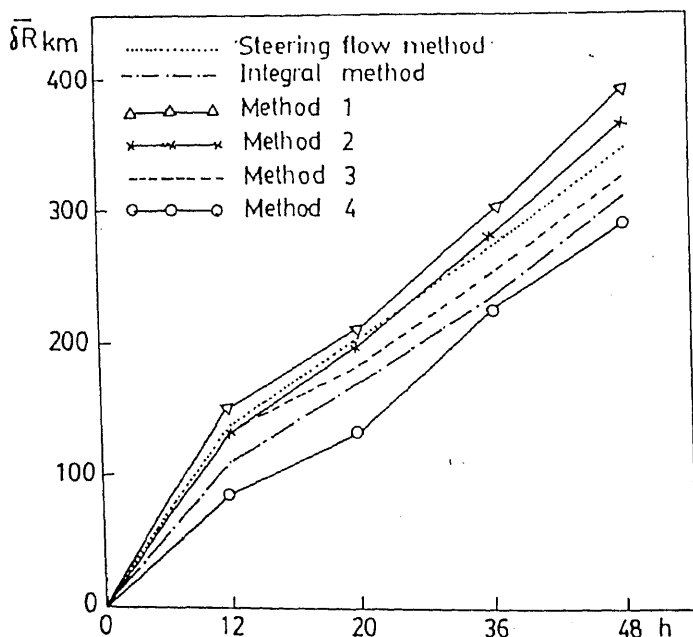


Figure 5. Performance (average magnitude of vector error) of specified objective methods in prediction of tracks of post-monsoon cyclones over the Bay of Bengal.

recurvature, which occurs mainly in the post-monsoon season, is quite a difficult task; but considering all these aspects, statistical method 4 is found to be appreciably better than others.

## 6. Conclusions

Based on the results presented in this study, the following general conclusions are drawn.

- (i) Although, two severe cyclonic storm seasons are generally observed in the Bay of Bengal, the primary one is the post-monsoon period. Also, it is during this season that the maximum number of complex tracks (i.e. recurved path of cyclonic storm motion) develop whose prediction is quite difficult.
- (ii) The dynamical approach based on the integral method is found to be the most attractive for predicting the motion of tropical cyclones, the complex multi-level PE models being the most suitable ones.
- (iii) In the present context, however, with lack of adequate meteorological data in and around the tropical cyclones and inaccuracies in parameterization of their physical processes, the integrated statistical approach (method 4) is found to be more satisfactory than the dynamical methods and other statistical methods discussed in this paper.

However, the predictive potential of dynamical models which are now greatly improved can be exploited with great effect with improved data coverage, generation of synthetic data, diabatic initialization and more accurate parameterization of physical processes.

## References

- Bansal R K, Datta R K 1974 A statistical method of forecasting the movement of cyclonic storms in the Bay of Bengal. *Indian J. Meteorol. Geophys.* 25: 391–397
- De Angelis D 1976 World of tropical cyclones – North Indian Ocean. *Mar. Weather Log.* 20: 191–194
- Gray W M 1968 Global view of the origin of tropical disturbances and storms. *Mon. Weather Rev.* 96: 669–700
- Hope J R, Neumann C J 1973 Operational use of the NHC-72 tropical cyclone forecast system: In Program of the 8th Technical Conference on Hurricanes and Tropical Meteorology. *Bull. Am. Meteorol. Soc.* 54: 147–188
- Hope J R, Neumann C J 1977 A survey of world-wide tropical cyclones prediction models. *Proceeding 11th American Meteorological Society Tech. Conference on Hurricane and Tropical Meteorology* (Am. Meteorol. Soc.) pp. 367–372
- India Met. Dept. 1979 Tracks of storms and depressions in the Bay of Bengal and the Arabian Sea, 1877–1970, India Meteorological Department
- Kasahara A 1959 A comparison between geostrophic and non-geostrophic numerical forecasts of hurricane movement with the barotropic steering model. *J. Meteorol.* 16: 377–384
- Kivganov A F, Mohanty U C 1978 On the physical statistical method for forecasting the movement of tropical storms in the Bay of Bengal. *Meteorol. Hydrol.* 5: 19–25
- Kivganov A F, Mohanty U C 1979a On the application of nonlinear regression analysis to forecast the movement of tropical cyclones in the Bay of Bengal. *Tr. Hydrometeorol. Centre, USSR* 125: 47–58

- Kivganov A F, Mohanty U C 1979b Hydro-dynamical method of forecasting the trajectories of tropical cyclones in the Indian seas. *Meteorol., Climatol. Hydrol.* 15: 9–15
- Minina L S, Niguen V N 1969 Tropical cyclones of Indian Ocean. *Tr. Hydrometeorol. Centre, USSR* 41: 29–50
- Mohanty U C 1978 *Tropical cyclones in the Bay of Bengal and objective methods for prediction of their movement*. Ph D thesis, Odessa Hydro-Meteorological Institute, USSR
- Mohanty U C 1979 Statistical method of forecasting tropical cyclones. *Meteorol. Climatol. Hydrol.* 15: 16–22
- Mohanty U C 1980 An objective method of forecasting the movement of tropical cyclones in the Bay of Bengal based on a number of statistical models. *WMO Symposium on Probabilities and Statistical Methods in Weather Forecasting*, Nice, France (Geneva: WMO) pp. 359–364
- Mohanty U C 1982 Some characteristics of dynamical initialization of mass and velocity fields in the lower latitudes. *Mausam* 33: 29–34
- Mohanty U C, Dube S K 1981 Statistical structure of the meteorological parameters over the Bay of Bengal during Monsoon-77 experiments. *Mausam* 32: 315–320
- Mohanty U C, Madan S C 1983 A numerical method for computing stream function and velocity potential. *Mausam* 34: 375–382
- Mohanty U C, Paliwal R K, Tyagi A, John A 1989 Evaluation of a multi-level primitive equations limited area model for short range prediction over Indian region. *Mausam* 40: 34–42
- Mohanty U C, Paliwal R K, Tyagi A, Sarin V B 1988 A suitable scheme of dynamic initialization for a multi-level primitive equations model in tropics. *Mausam* 39: 139–148
- Neumann C J 1972 An alternate to the Hurrell tropical cyclone forecast system – NOAA Tech. Memo. NWS SR-62
- Neumann C J, Hope J R, Miller B I 1972 Statistical method for combining synoptic and empirical prediction systems. NOAA Tech. Memo. NWS SR-63
- Neumann C J, Randrianarison E W 1976 Statistical prediction of tropical cyclone motion over the South-West Indian Ocean. *Mon. Weather Rev.* 104: 76–85
- Sanders F, Pike A C, Gaertner J P 1975 A barotropic model for operational prediction of tracks of tropical storms. *J. Appl. Meteorol.* 14: 256–280
- Sikka D R 1975 Forecasting the movement of tropical cyclones in the Indian seas by non-divergent barotropic model. *Indian J. Meteorol. Geophys.* 26: 323–325
- Sikka D R, Suryanarayan R 1972 Forecasting the movement of tropical storms/depressions in the Indian region by a computer oriented technique using climatology and persistence. *Indian J. Meteorol. Geophys.* 23: 35–40
- Simmons A J, Jarraud M 1984 The design and performance of the new ECMWF operational model. *ECMWF Seminar on Numerical Methods for Weather Prediction* (Reading: ECMWF) 2: 1–59
- Singh S S, Saha K 1976 Numerical experiments with a primitive equation barotropic model for prediction of movement of monsoon depressions and tropical cyclones. *J. Appl. Meteorol.* 15: 805–810
- Sitnikov E G 1968 Numerical experiments on prediction of track of tropical cyclones in the Republic of Cuba. *Tr. Hydrometeorol. Centre, USSR* 29: 21–32
- Slingo J M, Mohanty U C, Tiedtke M, Pearce R P 1988 Prediction of the 1979 summer monsoon onset with modified parameterization schemes. *Mon. Weather Rev.* 116: 328–346

# On the prediction of storm surges

P K DAS

A-59, Kailash Colony, New Delhi 110048, India

**Abstract.** The article provides a review of some recent work on the prediction of storm surges. Beginning with a historical account of major Bay of Bengal cyclones in the last two decades, it provides a description of surge prediction techniques. The assumptions that are made are discussed critically. This is followed by an account of the forcing terms that drive the surge, and the computational procedure for model prediction. A brief comparison is presented between model outputs and the observed peak surge for major storms. Finally, the interactions of the surge with the astronomical tide, and wind-generated waves are discussed. Although the main focus of the article is on storms in the Bay of Bengal there is a brief description of cyclones in the Arabian Sea. The article ends with suggestions for improving surge prediction in the years to come.

**Keywords.** Storm surges; Bay of Bengal cyclones; surge prediction techniques.

## 1. Introduction

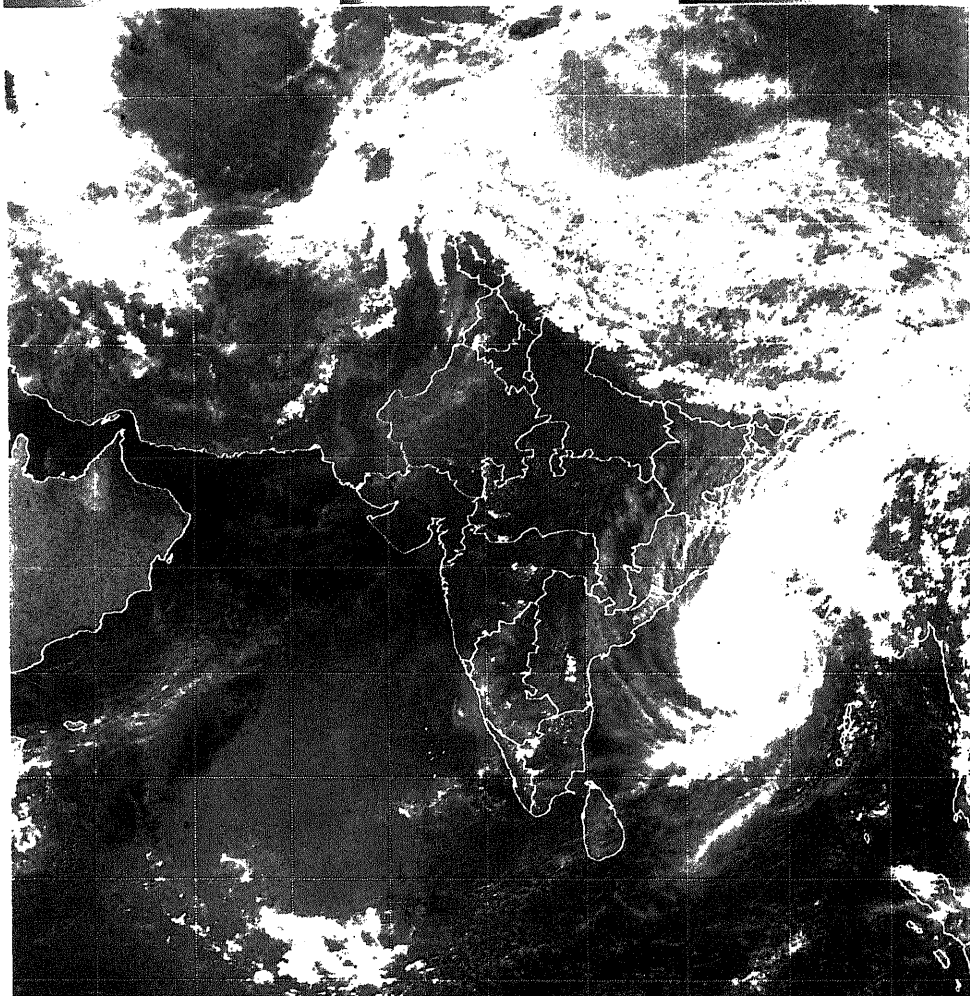
Storm surges belong to a genre of natural calamities that are classified as “windstorms”. They are induced by strong winds driving a large mass of sea water towards the coast. This leads to a sudden inexorable rise in sea level much above the normal tidal elevation. It causes much damage to life and property. Table 1 provides a list of six major storms in the Bay of Bengal that have struck India and Bangladesh in the last two decades (1970 to 1990).

It is worthwhile to note the decline in the number of fatal casualties on the Indian coast over the years. This is due to improvements in warning services and a closer network of coastal radars. The Indian Space Programme has also helped to reduce casualties. Of considerable importance is a system of disseminating cyclone warnings on fast satellite telecommunication links from Madras. This is the only one of its kind in the Third World.

Cloud imageries from the INSAT (Indian National Satellite) series of geostationary satellites have provided valuable inputs on storm features. The important features that can now be monitored on near real-time are : (a) the location of the storm centre, (b) the radius of an inner ring of calm winds round the centre, and (c) an estimate of the pressure deficit between the storm centre and the outer atmosphere. Item (b) is not always possible unless the images are sharp, but on many occasions an



INSATI INDIA MER 28-APR-91 09 00Z VIS VIS IR INSATID VIS ST VIS CR LINEAR  
IMDPS IMD NEW DELHI



Satellite view of the Bangladesh cyclone of April 29/30, 1993. The cyclone struck the Bangladesh coast between Hatia and Chittagong and caused 135,000 deaths. The sea level is reported to have risen by about 6 m above the mean. (Picture courtesy: India Meteorological Department.)

**Table 1.** Major tropical cyclones (1970–1990).

Year	Landfall	Country	No. of deaths (approx.)
1970	Chittagong	Bangladesh	300,000
1971	Paradip, Orissa	India	10,000
1977	Chirala, Andhra Pradesh	India	20,000
1982	Paradip, Orissa	India	250
1985	Hatia	Bangladesh	10,000
1990	Divi, Andhra Pradesh	India	250

approximate estimate is possible. Coastal radars also help to determine the radius of the eye.

Tropical cyclones are more frequent in the Bay of Bengal than in the Arabian Sea. The Bay cyclones, and their associated surges, are most frequent in the pre-monsoon months of April and May, and in the post-monsoon season from October to December. We will consequently focus on surges along the East Coast. The interested reader will find valuable information on storm surges, both in India and elsewhere, in a comprehensive review by Murty (1984).

## 2. Early studies

In early Indian work on surge prediction by empirical nomograms, Rao & Mazumdar (1966) used a nomogram that was designed earlier by Sverdrup & Munk (1947). A similar nomogram was used by Janardhan (1967) for estimating storm surges off Sagar Island at the northern end of the Bay of Bengal. Rao (1968) used the nomogram to delineate those sectors of the east and west coast of India which were vulnerable to dangerous surges. But, as these nomograms were based on empirical data the emphasis to-day is on predicting surges by numerical models.

Neumann & Mandal (1978) used a statistical regression equation to predict the track of a cyclone. By using the past positions of a storm at 12-hour intervals as predictors they found a linear regression equation for the 72-hour displacement of a storm. Interestingly, they found that a linear combination of only 4 predictors explained 65% of the variance in the future position of the storm. This shows that a large number of predictors are not always necessary for good prediction.

Chaudhury (1978) adopted a different approach. He used the method of images generated by a complex potential due to a vortex of constant vorticity. The region was bounded by two orthogonal planes representing the Indian and Bangladesh coasts. The apex of this region was located at the head of the Bay of Bengal. Although the method was mathematically simple and attractive, a doubtful assumption was made, which was that the atmosphere was an ideal fluid with no viscosity.

Many tropical cyclones in the Bay of Bengal begin as remnants of low pressure systems off the South China Sea. They intensify into tropical cyclones on entering the Bay of Bengal. Some, however, continue to move westwards and, on crossing the southern tip of the Indian Peninsula, emerge into the Arabian Sea. They then intensify again into tropical cyclones. Here they recurve and move towards the Indian coast. The most vulnerable sectors of the West Coast are (a) a part of the Konkan coast

north of 18°N, and (b) a sector off the Rann of Kutch from Dwarka to Karachi in Pakistan.

### 3. Surge models

#### 3.1 Basic equations

Surge prediction is built on the shallow water theory. This implies that the surge amplitude is small compared to the depth of the sea, which enables one to assume hydrostatic balance. The surge can thus be treated as a two-dimensional wave because the vertical component of motion is small in comparison with the horizontal components. A depth-averaged value of the current ( $\mathbf{V}$ ) is used in most conventional models but a few recent models are changing over to three dimensions. This provides a better representation of the vertical profile of  $\mathbf{V}$ .

The basic equations of a two-dimensional model are :

$$D\mathbf{V}/Dt + f\mathbf{K} \times \mathbf{V} + g\nabla(\zeta + p_a/g) + (\alpha/H)(\tau^S - \tau^B) + A_H \nabla^2 \mathbf{V} = 0, \quad (1)$$

$$\partial\zeta/\partial t + \nabla \cdot (H\mathbf{V}) = 0, \quad (2)$$

$$D/Dt \equiv \partial/\partial t + \mathbf{V} \cdot \nabla, \quad (3)$$

where

$\mathbf{V}$	= the depth-averaged current vector,
$\zeta$	= sea-level elevation,
$h$	= depth of the sea bed ( $Z = -h(x, y)$ ),
$H$	= total depth ( $\zeta + h$ ),
$f$	= Coriolis parameter ( $2\Omega \sin \phi$ ),
$p_a$	= atmospheric pressure,
$\alpha$	= specific volume,
$\tau^S, \tau^B$	= frictional stress at the surface and on the sea floor respectively,
$A_H$	= coefficient of eddy viscosity,
$\nabla$	= two-dimensional del operator,

and  $\mathbf{K}$  is the unit vector along the vertical axis of reference (figure 1).

By a scale analysis it is possible to show that the nonlinear terms in (1) and (2), and the effects of eddy viscosity, are small in comparison with the other terms as long as  $Z/H \ll 1$ , where  $Z$  is a characteristic amplitude of the surge (Charnock & Crease 1957); consequently, surge prediction models usually ignore nonlinear acceleration and eddy viscosity but the more recent models consider the complete equations.

#### 3.2 Forcing terms

**3.2a Atmospheric pressure:** A pressure deficit of 1 mb raises the sea level by 1 cm. This follows from hydrostatic balance. This rise in sea level is referred to as an inverted barometer effect. If we know the difference in pressure between the centre of a storm and the outer atmosphere then the inverted barometer effect provides a first guess to the rise in sea level. But, as the surge is mainly generated by strong winds, the inverted barometer effect is small in comparison with the wind stress. As we will see shortly, the pressure gradients in the vicinity of a storm determine the winds that drive the surge.

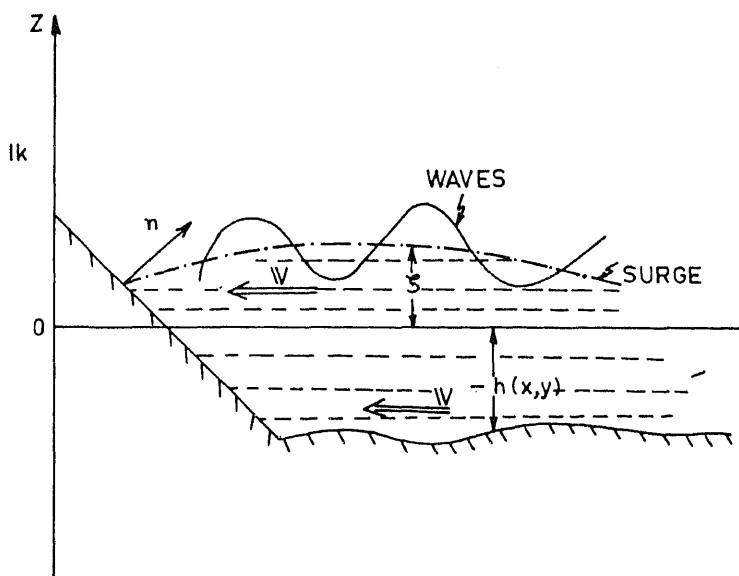


Figure 1. Schematic view of a surge.

3.2b *The wind stress and the sea bed friction:* The difference between the wind stress at the sea surface and the sea bed friction provides the main drive for the surge wave. The wind stress ( $\tau^s$ ) is related to the surface wind velocity ( $U$ ) at 10 m by

$$\tau^s = C_s \rho_a U |U|, \quad (4)$$

where  $\rho_a$  stands for the density of air and  $C_s$  is a surface drag coefficient.  $C_s$  varies with the wind speed and the stability of the atmosphere. Most modellers use a constant value of  $2.8 \times 10^{-3}$  for  $C_s$ , but empirical relations between  $C_s$  and  $U$  are also used.

The bottom stress ( $\tau^b$ ) is similarly related to the current vector ( $V$ ) by

$$\tau^b = C_b \rho V |V|, \quad (5)$$

where  $C_b$  is assigned a constant value of  $2.5 \times 10^{-3}$ .

Das (1972) assumed a steady Ekman spiral over the sea bed. For this current system

$$\tau^b = \alpha V, \quad (6)$$

where

$$\alpha = 0.035/h^2 s^{-1}. \quad (7)$$

The numerical constant (0.035) in the numerator of  $\alpha$  was determined empirically. The bottom stress is inversely proportional to the square of the depth.

There are uncertainties in this treatment. Ekman's spiral assumes a constant eddy stress coefficient, which is not realistic for the entire friction layer. Secondly, it is not clear whether the time taken to generate a boundary layer over the bottom of a shallow sea is comparable with the growth of the surge. In view of these uncertainties Flierl & Robinson (1972) omitted sea bed friction in their model for the Bangladesh cyclone of 1970.

Johns *et al* (1983) used another innovative approach. They used a closure scheme based on a balance equation for the turbulent kinetic energy ( $E$ ). The generation of  $E$  was balanced against (a) the rate at which  $E$  is extracted from the mean flow, (b) the vertical transport of  $E$ , and (c) the dissipation of  $E$  based on a similarity law. The scheme was used for a three-dimensional model of the Andhra Pradesh cyclone of 1977 (table 1). It was reported that there was not much difference in the final output between this scheme and the simpler equation (5). The determination of bottom stress remains one of the unsolved problems of surge prediction.

**3.2c Storm structure and winds:** The centre of a tropical cyclone is bounded by an inner ring of calm winds. This is the "eye" of the storm. The strongest winds are observed just beyond the eye-wall.

Earlier models used an idealised pressure profile to compute the associated winds. Thus, Das *et al* (1974) assumed the following profile

$$p_a = 1010 - \Delta p / [1 + (r/R)^2], \quad (8)$$

where  $\Delta p$  is the pressure deficit,  $r$  is the radial distance from the storm centre and  $R$  is the radius of the eye. Assuming a balance between the centrifugal force generated by rotating winds and the pressure gradient the following approximate expression was obtained for the maximum wind ( $V_m$ ).

$$V_m^2 = (13)^2 \times \Delta p, \quad (9)$$

where  $V_m$  is in knots and  $\Delta p$  is in millibars. An advantage of this method is that  $\Delta p$  can be estimated from real time satellite images of the storm. However, it assumes circular symmetry in the wind field.

Jelesnianski (1965) suggested another version. This is

$$V = V_m (r/R)^{3/2}; r \leq R, \quad (10a)$$

$$V = V_m (R/r)^{1/2}; r > R. \quad (10b)$$

Yet another empirical relation by Johns & Ali (1980) allows for a cut-off in wind strength beyond a distance  $R_1$  from the centre. This is not easy to determine for real-time prediction.

A certain degree of asymmetry was introduced by Jelesnianski & Taylor (1973) by the expression

$$V = V_m \times 2rR / (r^2 + R^2). \quad (11)$$

This is valid for a stationary storm. A precise definition of the wind field round a cyclone is another area of uncertainty, but there are indications to suggest that improvements in satellite and radar technology will help us to overcome this difficulty in the future.

#### 4. Coordinate transformation and computational procedure

A major difficulty in designing a model is to simulate an irregular coast. The coast is not a smooth mathematical curve which can be inserted in a model. Three different methods have been followed:

- (i) *Staircase models*: These models replace the coast by small horizontal and vertical segments in the form of a staircase (figure 2).
- (ii) *Continuous deformation*: Staircase models sometimes fail to capture sharp changes in coastal curvature. To overcome this difficulty Johns *et al* (1983) introduced the following transformation,

$$\xi = [x - b_1(y)]/b(y), \quad (12)$$

where

$$b(y) = b_2(y) - b_1(y), \quad b_1 \text{ \& } b_2 \text{ are constants,} \quad (13)$$

whence

$$\xi = 0, \quad x = b_1(y)$$

$$= 1, \quad x = b_2(y).$$

The independent variables are now  $\xi$ ,  $y$  and time ( $t$ ) for a two-dimensional model. This transformation has been extensively used by Dube *et al* (1981, 1982) for surges in the Bay of Bengal.

- (iii) *Conformal transformation*: Wanstrath (1976) introduced a conformal transformation to convert a curved domain into a rectangle. This avoids incorporating sharp changes in curvature. It has not been used much for the Indian coast, but a brief account has been reported by Mahadevan (1991).

At this stage it is difficult to assert which is best for the Indian coast because a detailed comparative study has not yet been made. Dube *et al* (1981, 1982) find that the staircase model tends to underestimate the peak surge by about 0.5 m in regions of rapid change in coastal curvature. This needs further study.

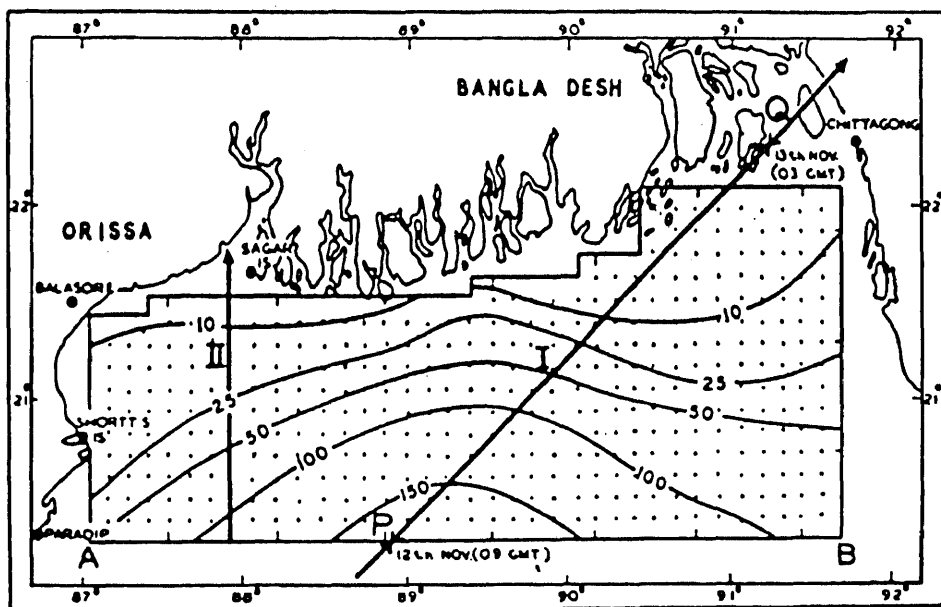


Figure 2. Grid for storms moving North-East (I) and North (II). Isopleths represent depth of the sea bed in metres (Das *et al* 1974).

## 4.2 Boundary conditions

The region covered by a prediction model has a land boundary in the form of an irregular coast, and an open sea boundary away from the coast. Conventional models assume a vertical wall along the coast so that the lateral boundary condition makes the normal component of  $\mathbf{V}$  vanish at the coast. We have

$$\mathbf{n} \cdot \mathbf{V} = 0, \quad (14)$$

where  $\mathbf{n}$  is a unit vector normal to the coast. A more realistic boundary condition which considers a sloping coast is,

$$\partial/\partial t(\zeta + h) + \mathbf{V} \cdot \nabla(\zeta + h) = 0, \quad (15)$$

but this has not been tried yet because of inadequate data on coastal topography.

Along the open sea a radiation boundary condition is usually employed. This allows outward propagation of energy from the region under consideration, but inhibits inward energy flow. Heaps (1969) expressed this formally by

$$v + (g/h)^{1/2} \zeta = 0, \quad (16)$$

where  $v$  is the meridional component of  $\mathbf{V}$ .

## 4.3 Computational procedure

The relevant equations are solved by considering their finite difference analogues. A staggered grid for the space variables and centred time differences are used. The grid is similar to Arakawa's (1966) energy conserving scheme, which conserves the kinetic energy in the region of interest, and prevents the growth of nonlinear instability. The stability criterion of Courant–Friedrichs–Levy (CFL) is used to determine the time steps. The criterion is,

$$(gH)^{1/2} \Delta t / \Delta x < 1, \quad (17)$$

where  $(gH)^{1/2}$  is the speed of gravity waves and  $\Delta t$ ,  $\Delta x$  are time and space increments. A nested grid, with higher resolution near the coast, was used by Das *et al* (1974). It provided better representation of the surge off the Orissa coast.

## 5. Model verification

The verification of model outputs is at present difficult because of an inadequate number of tide-gauges. On many occasions no tide-gauges exist near the landfall of a storm; consequently, the total sea level elevation is inferred from the tide-gauge nearest to landfall.

Interactions between the surge and the atmospheric tide are nonlinear. They could reduce the sea level elevation instead of raising it (Johns *et al* 1985).

In tables 2 and 3 we indicate the storm features, the tidal elevations and the observed maximum water level reported in different publications (Murty 1984; Dube 1992; Dube *et al* 1994). The results refer to the major storms mentioned earlier in table 1.

**Table 2.** Storm features.

Date	Landfall	Central pressure (mb)	Pressure deficit	Maximum wind speed (km/h)
13 Nov 1970	Chittagong, Bangladesh	940	70	222
30 Oct 1971	Paradip, Orissa	960	40	167
19 Nov 1977	Chirala, Andhra Pradesh	909	101	250
3/4 Jun 1982	Paradip, Orissa	950	50	216
24 May 1985	Hatia, Bangladesh	975	21	120
9 May 1990	Divi, Andhra Pradesh	920	80	230

Table 3 shows reasonable agreement despite uncertainties, between the computed surge and the observed surge + tide, if we superpose the tide on the computed surge. But, the verification programme will improve if more tide gauge records were available.

Based on model outputs nomograms have been prepared by Das *et al* (1974) for the surge ( $\zeta$ ) as a function of the pressure deficit ( $\Delta p$ ) and the speed of the storm ( $C$ ). We have

$$\zeta = a_0 \Delta p + a_1 (\Delta p)^2 + a_2 C, \quad (18)$$

where  $a_0$ ,  $a_1$  and  $a_2$  are constants. The nomograms were prepared for storms that move (a) northeast towards Bangladesh, (b) northwards towards the Hooghly estuary, and (c) northwest to the Orissa coast. Numerical values of the constants are available in the original publication and will not be reproduced here, but a typical nomogram is shown in figure 3. The nomograms are for pressure deficits less than 50 mb, but could be extended to higher values by (18), if required. A feature of the nomograms is a gradual approach to a constant surge amplitude for higher values of storm speed ( $C$ ). This is to be expected because the faster storms provide less time for a response from the sea.

Similar nomograms were prepared by Jelesnianski (1974) for a generalised coastline. It is referred to as SPLASH (Special Programme for Listing Amplitudes of Surges from Hurricanes). Ghosh (1977) has applied this nomogram for the Bay of Bengal cyclones. SPLASH was later replaced by SLOSH, which stands for "Sea Level over Land Surges from Hurricanes".

Each nomogram has its own limitation because each assumes a hypothetical storm with a specified structure. Consequently, in recent years the trend is to compute the

**Table 3.** Computed peak surge and observed surge + tide.

Date	Computed peak surge (m)	Tide (m)	Observed surge + tide (m)
13 Nov 1970	4.1	1.8	6.0 - 9.0
30 Oct 1971	3.5	0.9	4.0 - 5.0
19 Nov 1977	5.0	0.3	5.0 - 6.0
3/4 June 1982	3.5	1.3	> 3
24 May 1985	2.0	1.7	4.5(*)
9 May 1990	4.3	0.3	4.0 - 5.0

(\*) As reported by Bangladesh Meteorological Department



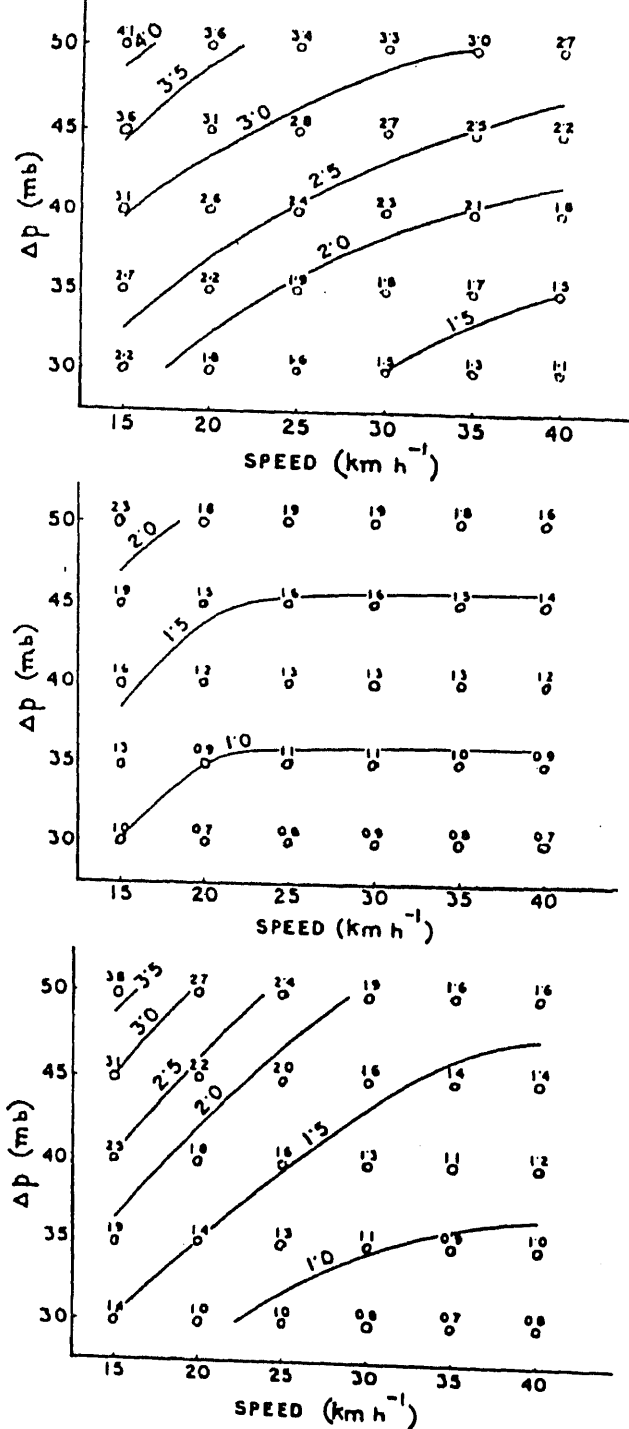


Figure 3. Nomograms from model generated storms: Sea level elevation ( $\zeta$ , m), storm intensity ( $\Delta p$ , mb) and speed ( $C$ , km/h); North-East (I) (a), Northward (II) (b), and North-West (III) (c) tracks. (a) and (b) correspond to tracks (I) and (II) of figure 2.

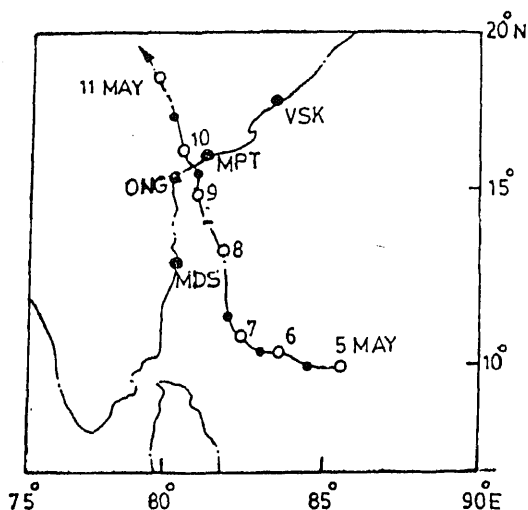


Figure 4. Path of the Andhra Pradesh cyclone of 1990.

surge for each storm on a small computer located at a forecasting centre. A standard procedure and a programme for this has been worked out by Dube (1992). This will benefit real-time surge prediction in the country. The prediction of the 1990 Andhra cyclone is illustrated in figures 4 and 5. This was done on a small computer.

## 6. Future developments and summary

There is considerable interest now in a coupled surge-tide-wave model. This is of much practical value because, as we can see, the total sea level elevation is the outcome of interactions between a surge, an astronomical tide and wind-driven waves. The period of the lunar semi-diurnal ( $M_2$ ) tide and the surge is of the same order (5–12 h). They are long gravity waves. But the period of a wind wave, which is a short gravity wave, is measured in seconds. We thus have a high frequency oscillation superposed on a low frequency oscillating system. There is little quantitative analysis of the interactions, or energy exchange, between these systems. The interactions will be most prominent in shallow water.

The other features which require further research are : (i) the Doppler shift in the frequency of waves on a current, (ii) an additional stress, known as the radiation stress, exerted by the waves on the surge, and (iii) the refraction of waves approaching a coast.

The principal conclusions may thus be summarised as follows.

- (i) There is need for better data on sea level elevation near the landfall of a storm. This could be achieved by a closer network of tide gauges, and by observations of sea level from space.
- (ii) Surge-prediction models may incorporate a sloping beach instead of assuming a vertical wall.
- (iii) Better data on the wind profile associated with a storm could improve prediction.
- (iv) Further research is needed for a more precise form of sea bed friction.
- (v) A coupled surge-tide-wave model will provide better estimates of total sea-level elevation.

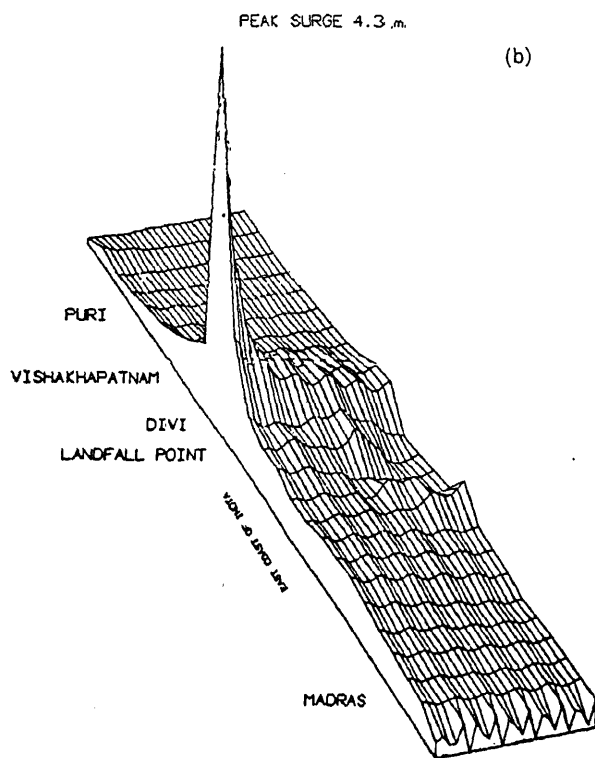
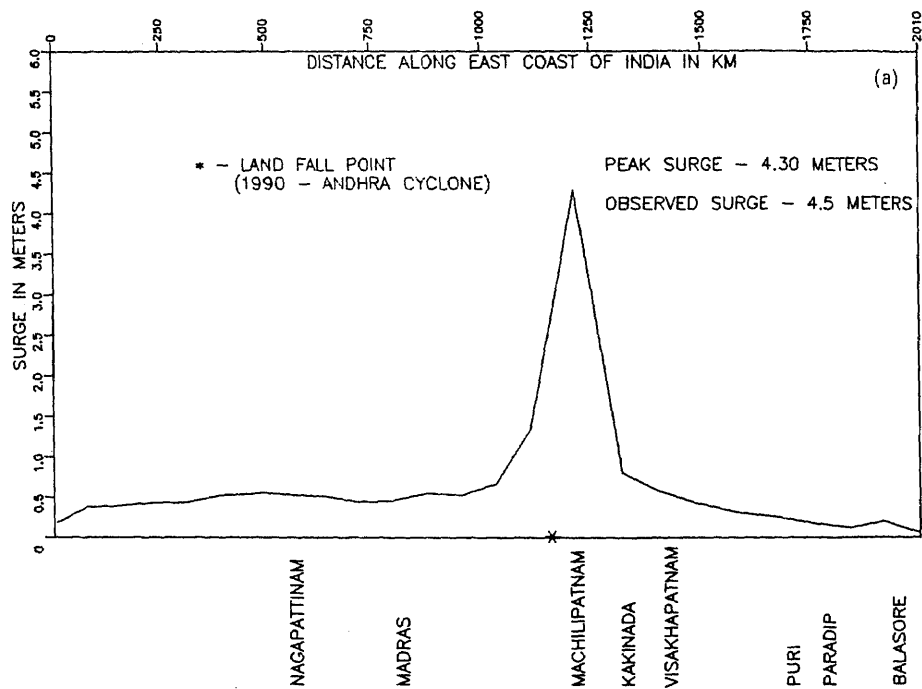


Figure 5. (a) Surge profile computed on near real-time, and (b) computer-generated 3-D surge profile for the Andhra Pradesh cyclone of 1990 (Dube *et al* 1994).

Dr G S Mandal of the India Meteorological Department kindly assisted me by providing information on recent cyclonic storms. I am indebted to him and Mr S D Gaur for their help.

## References

- Arakawa A 1966 Computational design for long-term numerical integration of the equations of atmospheric motion. *J. Comput. Phys.* 1: 119–143
- Chaudhury A M 1978 Rose petals for tropical cyclones. *Nucl. Sci. Appl.* A2: 1–7
- Charnock H, Crease J 1957 Recent advances in science: North Sea surges. *Sci. Prog.* 45: 494–511
- Das P K 1972 A prediction model for storm surges in the Bay of Bengal. *Nature (London)* 239: 211–213
- Das P K, Sinha M C, Balasubramanyam V 1974 Storm surges in the Bay of Bengal. *Q. J. R. Meteorol. Soc. (London)* 100: 437–449
- Dube S K 1992 Numerical modelling of storm surges in India and neighbourhood. *Proc. ICSU/WMO Seminar on Tropical Cyclone Disasters*, Beijing
- Dube S K, Rao A D, Sinha P C, Chittibabu P 1994 A real-time storm surge prediction system: An application to the east coast of India. *Proc. Indian Natl. Sci. Acad.* A60: 157–170
- Dube S K, Sinha P C, Rao A D 1981 The response of different wind stress forcings on the surge along the east coast India. *Mausam* 32: 315–320
- Dube S K, Sinha P C, Rao A D 1982 The effect of coastal geometry on the peak surge. *Mausam* 33: 445–450
- Flierl G, Robinson A 1972 Deadly surges in the Bay of Bengal, dynamics and storm tide tables. *Nature (London)* 239: 213–215
- Ghosh S K 1977 Prediction of storm surges on the east coast of India. *Indian J. Meteorol., Hydrol. Geophys.* 28: 157–168
- Heaps N 1969 A two-dimensional numerical sea model. *Philos. Trans. R. Soc. (London)* A265: 93–137
- Janardhan S 1967 Storm induced sea level changes at Saugor Island. *Indian J. Meteorol. Geophys.* 18: 205–212
- Jelesnianski C P 1965 A numerical calculation of storm tides induced by a tropical storm impinging on a continental shelf. *Mon. Weather Rev.* 93: 343–358
- Jelesnianski C P 1974 SPLASH NOAA Tech. Mem. NWS-TDL-52, NOAA, Washington DC
- Jelesnianski C P, Taylor A D 1973 A preliminary view of storm surges before and after storm modification. NOAA Tech. Memo. ERL-WMPO.3, NOAA, Washington DC
- Johns B, Ali A 1980 The numerical modelling of storm surges in the Bay of Bengal. *Q. J. R. Meteorol. Soc.* 106: 1–18
- Johns B, Rao A D, Dube S K, Sinha P C 1985 Numerical modelling of tide-surge interaction in the Bay of Bengal. *Philos. Trans. R. Soc. (London)* A313: 507–535
- Johns B, Sinha P C, Dube S K, Mohanty U C, Rao A D 1983 Simulation of storm surges using a three-dimensional numerical model: an application to the 1977 Andhra Cyclone. *Q. J. R. Meteorol. Soc.* 109: 211–224
- Mahadevan R 1991 Storm surges studies at Indian Institute of Technology, Madras – a review. *Proc. Seminar on Mathematical Modelling of the Ocean* (Bangalore: C-MMACS) p. 13a
- Murty T S 1984 *Storm surges, meteorological ocean tides* (Ottawa: Dept. Fisheries and Oceans)
- Neumann C J, Mandal G S 1978 Statistical prediction of storm motion over the Bay of Bengal and Arabian Sea. *Indian J. Meteorol., Hydrol. Geophys.* 29: 487–500
- Rao N S B 1968 *On some aspects of local and tropical storms in India*. Ph D thesis, Univ. of Jadavpur
- Rao N S B, Mazumdar S 1966 A technique for forecasting storm waves. *Indian J. Meteorol. Geophys.* 17: 333–346
- Sverdrup H U, Munk W H 1947 Wind, sea and swell: Theory of relations for forecasting. H F publication no. 601, US Navy
- Wanstrath J J 1976 Storm surge simulation in transformed coordinates, Vol. II: Program documentation. Tech. Rep. No. 76–3, US Army Corps of Engineers, VA



# Origin, incidence and impact of droughts over India and remedial measures for their mitigation

D A MOOLEY

Indian Institute of Tropical Meteorology, Pune 411 008, India

**Abstract.** The paper highlights the factors which tend to reduce the Indian monsoon rainfall substantially. These are: Eurasian snow accumulation during December–March, El Nino with warming phase of the eastern equatorial Pacific [0–10°S, 80°W–180°W], long periods of ‘break’ in the monsoon and poor westward penetration of monsoon low-pressure systems.

Utilising the joint criteria based on specified threshold values of the percentage departure from normal for area-averaged Indian monsoon rainfall and of the percentage Indian area under deficient monsoon rainfall, the years of all-India drought have been identified. Enhancing these criteria suitably, years of severe and phenomenal droughts have also been identified. India experienced 18 droughts during 1871–1990, of which 10 were severe and 5 were phenomenal. While the periods 1901–20 and 1961–80 had the highest frequency of drought, 1921–40 had no drought at all. Most of the severe, and all the phenomenal droughts, occurred in the El-Nino-cum-warming phase which, therefore, needs to be carefully monitored.

The remedial measures which can be actively pursued are effecting improvements in agriculture and in ruraleconomy, according high priority to population control, initiating suitable insurance schemes, generating confidence amongst the affected people and making adequate financial provision to render relief effectively and quickly to drought-hit communities.

**Keywords.** Droughts; El-Nino; drought remedial measures.

## 1. Introduction

The water required for different activities in India is met mostly from rainwater provided by the summer monsoon, because a large part of the country gets most of its annual rainfall during the four monsoon months, June to September. In view of the explosive growth of population in India during the last 15 to 20 years, the demand for water has increased substantially, and the Indian economy has become vulnerable to rainfall deficiency.

Due to the high concentration of rainfall in the monsoon season, Indian droughts generally result from failure of the summer monsoon. There are three types of drought: meteorological, hydrological and agricultural. While considering meteorological

drought, the degree of dryness is considered in comparison to the normal rainfall. A definition of meteorological drought can be based on some specified threshold rainfall, or its anomaly. Hydrological drought follows meteorological drought and is reflected in reduced stream flow, reduced reservoir/lake levels, and reduced groundwater flow. Agricultural drought results from soil moisture deficiency. The deficiency depends on the water needs of plants and the availability of water. Agricultural drought is highly complex as the water requirements of crops depend on different stages in their growth. It varies from one crop to another.

Here, we shall consider only meteorological droughts.

## 2. Factors leading to deficient monsoon rainfall

### 2.1 Global factors

**2.1a Eurasian snow cover during December to March:** If the snow cover is much above normal, a part of the incoming solar radiation during April and May is used up in melting the excess snow and the remainder is used to heat the South-Asian land mass and the overlying atmosphere. Consequently, seasonal heating would be deficient and the monsoon would not develop well, resulting in deficient rain. Thus, a high snow cover would result in low monsoon rainfall. An inverse significant relationship is generally found for the period 1967–1979 between the Indian monsoon rainfall and the Eurasian snow cover from December to March as obtained from satellite data (Hahn & Shukla 1976; Dickson 1984). But, the inclusion of later data shows some deterioration in the relationship (Mooley & Shukla 1987, pp. 26–59). The amount of snow during December to March may serve as a better parameter. Satellites generally provide information on the area of snow cover. This parameter, however, needs to be monitored in view of its potential for adversely influencing the subsequent Indian monsoon rainfall. Satellite snow cover data are obtained from the US Climate Analysis Center.

**2.1b El Nino, sea surface temperature of eastern equatorial Pacific and southern oscillation:** The El Nino originates off the Peru–Ecuador coast (0–12°S) in South America around Christmas time. El Nino is a Spanish word which means ‘the Child’. The origin is thus associated with the birth of Christ. This phenomenon is an occasional warming off the Peru–Ecuador coast (0–12°S) and results in a rise of the sea surface temperature (SST) by about 2°C or more above normal. It usually lasts for about a year attaining a maximum SST off the Peru–Ecuador coast around June. Quinn *et al* (1978) and Rasmusson & Carpenter (1983) have tabulated El Ninos after a careful scrutiny of SST data. During 1871–1990 there were 29 El Ninos. In many El Nino years, the warming off the coast spread westwards on account of the southeast trade winds of the Southern Hemisphere.

The area-averaged seasonal SST anomaly for the large Pacific area, 0–10°S; 80–180°W (eastern equatorial Pacific) was studied by Mooley & Paolino (1989) for the period 1871–1984, using the data collected by Angell (1981) and later updated by him. Progressive warming over 3 or more seasons has been identified as a “warming phase” by them by using specific criteria. The specific criteria are: (i) total warming being 1°C or more, leading to attainment of an anomaly of 1°C or more, (ii) if the anomaly is less than 1°C, then the SST increase should be  $(1 + y)^\circ\text{C}$  or more, where

$y$  is the amount by which the anomaly attained falls short of  $1^{\circ}\text{C}$ . Thus, a smaller anomaly is compensated by a larger rise in SST. For example, if the anomaly is  $0.3^{\circ}\text{C}$ , then the SST rise should be  $1.7^{\circ}\text{C}$  or more. With these criteria, they identified 27 warming-phase years. During 1985–90, there was one such year, namely, 1987. Thus, there are 28 warming-phase years during 1871–1990. The mean SST anomaly in warming-phase years increased from the Sept–Nov season of the preceding year to the Dec–Feb season of the year following a warming-phase year. They found that the warming phase exerts an adverse influence on monsoon rainfall. This influence is due to displacement of the equatorial vertical Walker circulation cell and its interaction with monsoon circulation. On examination of El-Nino and warming-phase years, it is found that a large number of years are El-Nino-cum-warming-phase years. The period 1871–1990 is divided into four sets of years. The first set consists of years labelled as El-Nino-cum-warming-phase years. The second consists of warming-phase but not El-Nino years and are labelled as only warming-phase years. The third consists of El-Nino but not warming-phase years and are labelled as only El-Nino years. The fourth consists of years which are neither warming phase nor El Nino. The main features of area-averaged Indian monsoon rainfall departures in these sets of years are given in table 1. This table also gives the rank (within and including 10), when the years are ranked on the basis of increasing departure of Indian monsoon rainfall. The main features of table 1 are as follows.

- (i) *El-Nino-cum-warming-phase* – This set consisting of 19 years exerts a strong adverse influence on the Indian monsoon rainfall. The mean rainfall departure from normal for this set is  $-13.2\%$  with a range from  $-29.0\%$  to  $0\%$ . In this set, there are 11 years with percentage rainfall departure  $< -10.0\%$ . This set contains years with ranks 1 to 6 and 8. In all the years of this set, warming off Peru and Ecuador due to El Niño spreads to the whole of the eastern equatorial Pacific, and the progressive rise of SST over eastern equatorial Pacific satisfies the criteria for the warming-phase year.
- (ii) *Only warming phase* – This set includes 9 years. Some adverse influence on Indian monsoon rainfall is observed. The mean percentage rainfall anomaly is  $-4.8$  with a range from  $-12.0$  to  $+2$ . There are 2 years with rainfall anomaly  $< -10\%$ .
- (iii) *Only El Nino* – This includes 10 years. There appears to be no influence of an El Nino in these years on the Indian monsoon rainfall.
- (iv) *Neither El Nino nor warming phase* – This set consists of 82 years and is a mixed set with rainfall departure ranging from  $-17\%$  to  $+19\%$ . This set contains low rainfall years with ranks 7, 9 and 10. The low rainfall in the 8 years of this set is due to factors other than El Nino-warming phase or only the warming phase.

We find that while warming-phase years have a tendency to reduce Indian monsoon rainfall, it is the El-Nino-cum-warming-phase years that have a high or strong tendency to reduce the rainfall. Comparing the ranks in the El-Nino-cum-warming-phase set with those in the set of years with neither El Nino nor warming phase, we infer that an El Nino with warming phase has much higher potential for high deficiency of Indian monsoon rainfall.

In view of this situation, the El-Nino-cum-warming-phase of the eastern equatorial Pacific should be carefully monitored. The El Nino is being monitored at the US Climate Analysis Center. A watch on the warming of the Pacific region is also needed.

The Southern Oscillation is a see-saw between the atmosphere over the eastern Pacific and that over the western Pacific and the Indian Ocean. When the pressure



**Table 1.** Percentage departure from mean of Indian monsoon rainfall in (i) El-Nino-cum-warming-phase years (ii) warming-phase years only, (iii) El-Nino years only, (iv) neither El-Nino nor warming-phase years. Period covered: 1871-1990.

	El-Nino-cum-warming- phase years	El-Nino years only	Warming-phase years only	Neither El-Nino nor warming-phase years
No. of years	19	10	9	82
Mean departure	-13.2%	+0.2%	-4.8%	—
Range of departure	-29% to 0%	-7% to +9%	-12% to +2%	-17% to +19%
No. of years with departure $\leq -10\%$	11	0	2	8
Years with rank upto ten in order of increasing departure	1877(1), 1899(2), 1918(3), 1972(4), 1987(5), 1965(6), 1905(8)	None	None	1979(7), 1920(9), 1901(10)

over the eastern Pacific is high that over the western Pacific and the Indian Ocean is low, and vice versa. The pressure at Darwin is often used as a measure of the Southern Oscillation Index. The Southern Oscillation and the El Nino are closely linked with each other, and the two together are referred to as the ENSO phenomenon. Sikka (1980) has brought out large-scale fluctuations in Indian monsoon rainfall in response to planetary scale features of the wind circulation of the atmosphere.

A stable and significant relationship is found between the SST tendency from the Dec–Feb to the Mar–May season, with the preceding as well as the following Indian monsoon rainfall. However, a stable and significant relationship is also found between the Southern Oscillation tendency and the following Indian monsoon rainfall over the period 1942–90 (Mooley & Munot 1993).

## 2.2 *Regional factors*

**2.2a *Variations in the intensity and the location of the monsoon trough over India:*** A monsoon trough normally extends from Northeast Rajasthan to the Northwest Bay. It is a semi-permanent feature of the season. When the trough is active and is located in the normal position, the rainfall over the country is generally good and well-distributed. When the trough is weak, the rainfall is generally much less. However, when the trough shifts north to the Himalayan foothills a marked change occurs. Little rain falls over most of the country but torrential rain occurs in the sub-Himalayan region. Such situations are known as 'breaks' in the monsoon. A large number of days of 'breaks' reduces the rainfall appreciably.

**2.2b *Behaviour of westward moving low pressure systems:*** Lows and depressions form over the north Bay of Bengal and its neighbourhood. They generally move West or Northwest. In years of poor monsoons, these systems rarely move West of 80°E. They are either dissipated or recurve North/Northeast before reaching this longitude. In years of good monsoon, they travel further West upto 70°–75°E (Mooley & Shukla 1989). It has been found that the total westward displacement of the low pressure systems during the whole season is significantly related to the monsoon rainfall.

## 3. **All-India drought**

### 3.1 *Criteria for all-India drought and identification of drought years*

Many workers have studied the incidence of drought over India, using different criteria. Most of them have followed the criterion that drought is an annual/monsoon season rainfall deficiency of 25% or more (India Met. Dept. 1971; Govt. of India 1976; Koteswaram 1976).

Every year, some part of the country may experience drought. But here we consider droughts affecting a sizeable portion of India (termed All-India drought). A drought of local nature, that is, affecting a few meteorological sub-divisions or a state will not be considered here.

Kharif food production of India has a very good correlation coefficient (0.88) with area-averaged Indian monsoon rainfall for the period 1966–88 (Parthasarathy *et al* 1992). One of the notable effects of drought is a fall in foodgrain production. Hence

a criterion for drought based on area-averaged Indian monsoon rainfall would be reasonable. The criterion adopted is that the area-averaged Indian monsoon rainfall equals, or is less than the mean minus one standard deviation (SD). Since SD is about 10% of the mean for the area-averaged Indian monsoon rainfall, this criterion actually means that the Indian monsoon rainfall departure is equal to or less than  $-10\%$  of the mean. However, this criterion alone is not sufficient for identifying drought since adverse effects also depend on the rainfall deficiency over the different meteorological sub-divisions. Therefore, a second criterion which depends on the deficient rainfall of the sub-divisions has to be applied additionally. The effect of deficient rainfall over the different meteorological sub-divisions in a year is best integrated by considering the total Indian area under deficient monsoon rainfall. The second criterion is based on this area expressed as a percentage of the total Indian area. Now a sub-division with deficient rainfall is defined as one whose monsoon rainfall is less than or equal to mean minus SD. For each year, the sub-divisions with deficient monsoon rainfall are identified and their total area is calculated and is then expressed as a percentage of the Indian area. The percentage of deficient rainfall area criterion additionally adopted for identifying all-India drought is this deficient rainfall area  $\geq (\text{mean} + \text{SD})$ , i.e.  $\geq 31.4\%$ , since the mean is  $15.1\%$  and SD,  $16.3\%$ . Values of percentage area under deficient rain for 1871–1990, their mean and SD as tabulated by Parthasarathy *et al* (1992) have been utilized. Adopting these two criteria jointly, All-India droughts during the period 1871–1990 have been identified (table 2). It can be seen that there

**Table 2.** Years of all-India drought along with percentage departure of Indian monsoon rainfall, and country's percentage area under deficient rainfall [drought years based on joint criteria: percentage rainfall departure  $\leq -10.0$ ; percentage area under deficient rain  $\geq (\text{mean} + \text{SD})$  i.e.  $\geq (15.1 + 16.3)$  i.e.  $\geq 31.4$ ].

Year	Rainfall departure (%)	Area under deficient rainfall (%)
1873	-11.4	40.9
1877*	-29.1	66.8
1899*	-26.2	83.0
1901	-15.6	45.4
1904 <sup>+</sup>	-12.0	38.9
1905*	-16.1	40.1
1911*	-14.0	43.5
1918*	-23.9	68.2
1920	-15.8	36.5
1941*	-14.4	41.8
1951*	-13.5	36.3
1965*	-17.0	43.0
1966	-13.7	31.3 <sup>a</sup>
1972*	-23.3	49.5
1974	-12.3	34.1
1979	-16.9	49.2
1982*	-13.7	46.4
1987*	-19.3	64.3
Mean	-17.1	47.2

<sup>a</sup> Area criteria marginally satisfied

Note: \* El-Nino-cum-warming-phase year; <sup>+</sup> warming-phase year only

are 18 All-India droughts in the whole period giving an average of one drought in 6 or 7 years. The 20-year period with highest frequency of 6 is 1901–20 and the next 20-year period is 1961–80 with 5 droughts. The 40-year period 1921–60 between these two 20-year periods has a frequency of only 2 droughts. The 20-year period 1921–40 had no drought at all. Out of 18 droughts, 11 occurred in the years of El-Nino-cum-warming phase, one (1904) occurred in the year of the warming phase only, and none occurred in the years of El Nino alone. There are 2 cases of successive drought years, 1904–05 and 1965–66.

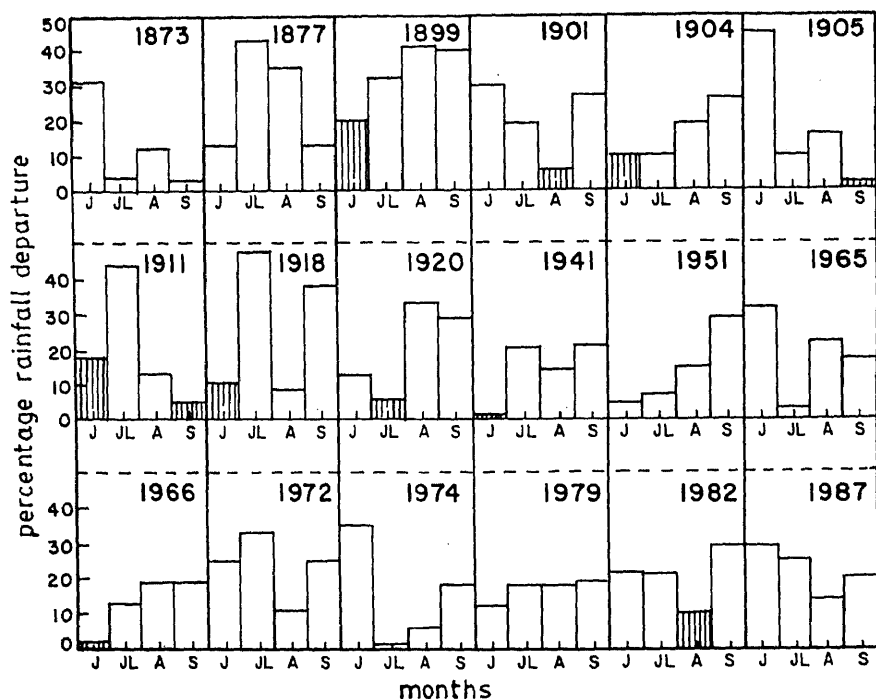
We would like to classify severe and phenomenal all-India droughts. For this purpose, we have to utilise higher levels of the two criteria i.e. more stringent criteria. For severe drought, the levels of the criteria are enhanced by 50%. These, on enhancement, become, percentage departure of the area-averaged Indian monsoon rainfall  $\leq -15\%$ , and the country's area under deficient monsoon rainfall  $\geq (\text{mean} + 1.5 \text{ SD})$ , i.e.  $\geq 39.5\%$ . Utilizing these joint criteria with these levels, severe drought years have been identified. These are 1877, 1899, 1901, 1905, 1918, 1941, 1965, 1972, 1979 and 1987. There have been 10 severe droughts, giving an average of one every 12 years. Severe droughts in succession have not occurred during the entire period. All except two (in 1901 and 1979) have occurred in years of El-Nino-cum-warming phase. India experienced 2 severe droughts in the decades 1901–10 and 1971–80, none in the decades 1881–90, 1921–30, 1931–40 and 1951–60 and one severe drought in each of the remaining six decades. In the years of these severe droughts the mean rainfall departure was  $-18.9\%$  and the country's mean area under deficient rain was  $53.8\%$  (more than half the country).

Utilizing still more stringent joint criteria, viz. rainfall departures  $\leq -20\%$  and country's area under deficient rain  $\geq (\text{mean} + 2.0 \text{ SD})$  i.e.  $\geq 47.7\%$ , for phenomenal drought, phenomenal droughts during 1871–1990 have been identified. These are 1877, 1899, 1918, 1972 and 1987. There are 5 phenomenal droughts, giving an average of one in 24 years. All the five droughts occurred only in years of El-Nino-cum-warming-phase. Initially, phenomenal droughts occurred after almost 20 years. There was a long gap of 53 years after 1918. Mean rainfall departure in phenomenal droughts is  $-24.4\%$  and country's mean area under deficient rainfall is  $64.4\%$ . Considering the rainfall departure and the country's area under deficient rain, the phenomenal droughts of 1877 and 1899 are the severest, both being approximately of the same severity. The 1918 drought though comparable to these two is slightly lesser in severity. Thereafter, during the last 75 years, phenomenal droughts have been relatively smaller in intensity. If a drought of the intensity of 1877 or 1899 were to occur now or in the future, the devastation would be far greater in view of the much higher level of population. There is no reason to believe that a drought of this intensity may not occur in India in future. And if it occurs, there is a high probability of its occurrence in a year with a severe El-Nino-cum-warming-phase.

### 3.2 *Monthly percentage Indian rainfall departure in drought years*

The monthly percentage rainfall departures for June, July, August and September are shown in figure 1. The main features of these are given below.

- (i) In June, 4 years recorded positive departure  $> 10\%$ , the highest being  $20\%$  in 1899 and the next highest being  $18\%$  in 1911. The lowest departure for June was  $-45\%$  in 1905. The mean for June is  $-13\%$ .



**Figure 1.** Percentage Indian monthly rainfall departure from normal in drought years – ▨ positive □ negative. Normal rainfall (mm): June 163, July 275, August 243, September 171, mean departure for drought years (%) June – 13, July – 19, August – 16, September – 20.

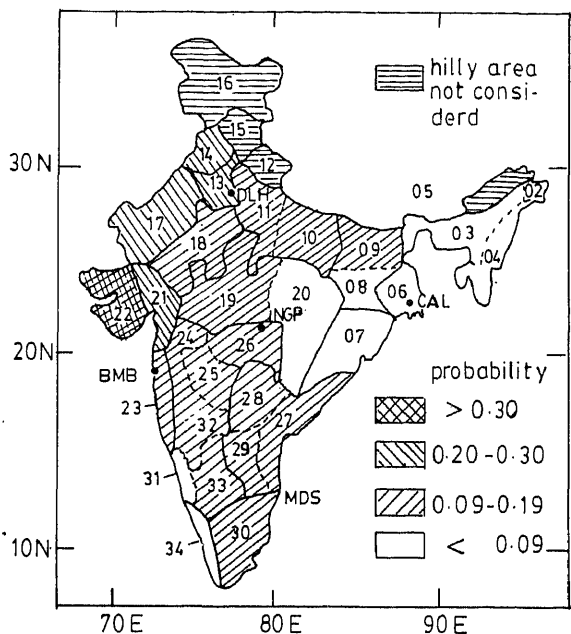
- (ii) In July, departure varied from + 6% in 1920 to – 48% in 1918, with a mean of – 19%.
- (iii) In August, the departure varied from + 5% in 1901 to – 41% in 1899, with a mean of – 16%.
- (iv) In September, the departure varied from + 5% in 1911 to – 40% in 1899, with a mean of – 20%.

In years of severe droughts, the monthly rainfall departure is  $< -20\%$  in at least two months and is negative in at least three months.

In years of phenomenal droughts the mean monthly rainfall departure is  $-20\%$  averaged over four monsoon months, and is  $< -25\%$  in July, suggesting strong contribution by this month. In 1877 and 1899, July and August, the core monsoon months, contributed strongly to the deficiency.

#### 4. Incidence of rainfall deficiency over different meteorological sub-divisions of India

Here the incidence of deficiency of  $> 25\%$  from normal over various sub-divisions has been considered. The probability of incidence of this deficiency has been computed for each of the sub-divisions on the basis of data for 1871–1990. This is shown in figure 2. The probability exceeds 0.30 for Saurashtra–Kutch, the westernmost sub-division of India. Harayana, Punjab, West Rajasthan and Gujarat sub-divisions



- |                              |                           |
|------------------------------|---------------------------|
| 2 Arunachal Pradesh          | 19 West Madhya Pradesh    |
| 3 North Assam                | 20 East Madhya Pradesh    |
| 4 South Assam                | 21 Gujarat                |
| 5 Sub-Himalayan West Bengal  | 22 Saurashtra & Kutch     |
| 6 Gangetic West Bengal       | 23 Konkan                 |
| 7 Orissa                     | 24 Madhya Maharashtra     |
| 8 Bihar Plateau              | 25 Marathwada             |
| 9 Bihar Plains               | 26 Vidarbha               |
| 10 East Uttar Pradesh        | 27 Coastal Andhra Pradesh |
| 11 West Uttar Pradesh Plains | 28 Telengana              |
| 12 West Uttar Pradesh Hills  | 29 Rayalseema             |
| 13 Haryana                   | 30 Tamil Nadu             |
| 14 Punjab                    | 31 Coastal Karnataka      |
| 15 Himachal Pradesh          | 32 North Karnataka        |
| 16 Jammu & Kashmir           | 33 South Karnataka        |
| 17 West Rajasthan            | 34 Kerala                 |
| 18 East Rajasthan            |                           |

**Figure 2.** Empirical probability of monsoon rainfall deficiency ( $\geq 25\%$ ) over different meteorological sub-divisions of India (1871-1990).

have a probability of 0.20 to 0.29. Some of the eastern sub-divisions and the southern part of the West Coast have a low probability of  $< 0.10$ . The remaining sub-divisions have a probability of 0.10 to 0.19. In general, the probability over the country decreases from west to east.

## 5. Impact of drought

Different aspects of the impact of drought over India are briefly indicated below. Due to the present explosive growth in population the impact would now be more severe.

### 5.1 *Economic*

- (i) Losses in production of kharif crops, dairy and livestock, fisheries and hydro-electric power.
- (ii) Loss of income to farmers, farm labourers (loss of wages), Central and State Governments (revenue loss due to reduced tax base).
- (iii) Shortages of (a) water for drinking, irrigation, hydro-electric power, and industry, (b) foodgrains.
- (iv) Increase in price of foodgrains (by as much as 50% to 100% in drought-hit areas).
- (v) Increase in rural indebtedness owing to money lenders exploiting the situation.
- (vi) Incidence of epidemics. Subsistence at low levels of food-intake resulting from reduced purchasing power leads to low vitality and higher vulnerability to epidemics.

### 5.2 *Environmental*

- (i) Damage to animal species due to lack of feed/water and due to diseases.
- (ii) Wind erosion of soils.
- (iii) Effect on air and water quality (due to dust and pollutants).

### 5.3 *Social*

- (i) Exploitation of shortages by fraudulent traders.
- (ii) Food riots. Those getting neither employment nor food resort to looting foodgrains to satisfy hunger.
- (iii) Spread of epidemics and resulting loss of life.

## 6. **Remedial measures**

Remedial measures are indicated below. Some of these are already being pursued. However, to keep pace with the alarming growth of India's population, these measures will have to be pursued vigorously. Alternatively, adequate brake on population growth has to be initiated.

- (I) Improvements in agriculture—
  - (a) increase in irrigation works/wells/tanks;
  - (b) sowing high-yielding and drought-resistant crop varieties;
  - (c) using better fertilizers, and more effective insecticides and pesticides;
  - (d) stepping up dry land farming.
- (II) Strengthening of rural economy.
- (III) Building up food reserves in rural areas also.
- (IV) Improving rail/road communication in drought-prone areas.
- (V) Adequate financial provisions –

Creation of a 'Natural Disaster Relief Fund' would go a long way in providing a cushion to affected communities in the aftermath of natural disasters which hit the country almost every year. A fund of this kind could be easily established through creative management of emergency allocations and donations received in the wake of natural disasters, and its existence would stimulate the evolution of longer lasting

protection measures that would provide long-term safeguards to vulnerable communities.

(VI) Insurance—

- (a) It would be advisable to introduce attractive insurance schemes to bring a large section of farmers under crop insurance;
- (b) The feasibility of a scheme to cover loss of wages to farm labourers may be explored by the General Insurance Corporation of India;
- (c) The feasibility of a Cattle Insurance Scheme may be explored by the Life Insurance Corporation of India.

(VII) Suggestions for efficient drought management and for maintaining continuity of stated drought policy and of action—

- (a) constitution of a permanent National Disaster Relief Commission with a committee for each type of disaster;
- (b) statements may be made on Drought Policy and Objectives of Drought Plan;
- (c) implementation of a standard drought plan for relief to drought-hit areas should be initiated;
- (d) Setting up of a monitoring unit to monitor drought conditions. For timely action in mitigating the ensuing hardship, complete coordination between the monitoring unit and all those executive officers concerned with mitigation work is essential. Timely action will inspire confidence and satisfaction among the affected population, and would make the mitigation process smooth and fast;
- (e) establish a committee to identify research needs for assessment of drought and its impact;
- (f) develop a training programme for personnel working for drought relief, for implementation as required;
- (g) to evolve evaluation procedures for relief rendered and hardship mitigation *vis-a-vis* the drought plan;
- (h) evaluation of each drought and preparation of a report on the same.

## 7. Concluding remarks

- (i) The large Eurasian snow accumulation during December–March, the El Nino with warming phase of eastern equatorial Pacific, the long periods of ‘break’ in monsoon/weak monsoon trough, and poor westward penetration of monsoon lows, which have the potential to reduce monsoon rainfall substantially, need to be carefully monitored.
- (ii) During 1871–1990, India experienced 18 droughts of which 10 were severe and 5 were phenomenal. While the periods 1901–20 and 1961–80 had the highest drought incidences, the period 1921–40 experienced no drought at all. All phenomenal droughts and most of the severe droughts occurred only in the El-Nino-cum-warming-phase of the eastern equatorial Pacific.
- (iii) Drought-proneness generally increases westwards.
- (iv) Improvements in agriculture, strengthening of the rural economy, building up of food reserves in rural areas, some high priority measures to control population, timely action to create confidence and satisfaction amongst the affected people, suitable schemes of insurance, adequate financial provision through the creation of a permanent Natural Disaster Relief Fund, as well as of a permanent National



Commission for Natural Disaster Relief are some of the remedial measures suggested.

The author is grateful to Dr G B Pant for providing facilities, to Dr J K Angell for the sea surface temperature data of the eastern equatorial Pacific, to Dr H N Srivastava for the monthly rainfall data of the raingauge stations, to Dr B Parthasarathy for the area-averaged rainfall of India, and to Ms S P Lakade, Ms J V Revadekar and Ms S S Nandargi for help in various forms.

## References

- Angell J K 1981 Comparison of variation of atmospheric quantities with sea surface temperature variations in the equatorial eastern Pacific. *Mon. Weather Rev.* 109: 230–243
- Dickson R 1984 Eurasian snow-cover versus Indian rainfall – an extension of Hahn–Shukla results. *J. Climatol. Appl. Meteor.* 23: 171–173
- Govt. of India 1976 Climate and agriculture. Part IV of the report of the National Commission on Agriculture, Ministry of Agriculture & Irrigation, New Delhi
- Hahn D, Shukla J 1976 An apparent relationship between Eurasian snow-cover and Indian monsoon rainfall. *J. Atmos. Sci.* 33: 2461–2463
- India Met. Dept. 1971 Rainfall and droughts in India. Report of the Drought Research Unit, India Meteorological Department, Pune
- Koteswaram P 1976 Climatological studies of droughts in Asiatic monsoon area particularly India. *Proc. Indian Natl. Sci. Acad.* 54: 1–14
- Mooley D A, Munot A A 1993 Variation in the relationship of the Indian summer monsoon with global factors. *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* 102: 89–104
- Mooley D A, Paolini D A 1989 The response of the Indian monsoon associated with the changes in sea surface temperature over eastern south equatorial Pacific. *Mausam* 40: 369–380
- Mooley D A, Shukla J 1987 Variability & forecasting of the summer monsoon rainfall over India. In *Monsoon meteorology* (eds) C P Chang, T N Krishnamurti (Oxford: University Press)
- Mooley D A, Shukla J 1989 Main features of the westward-moving low pressure systems which form over Indian region during the monsoon season and their relationship with the monsoon rainfall. *Mausam* 40: 137–152
- Parthasarathy B, Rupa Kumar, Kothawale D R 1992a Indian summer monsoon rainfall indices: 1871–1990. *Meteorol. Mag.* 121: 174–186
- Parthasarathy B, Rupa Kumar K, Munot A A 1992b Forecasting of rainy season foodgrain production based on monsoon rainfall. *Indian J. Agri. Sci.* 62: 1–8
- Quinn W H, Zoff D O, Short K S, Kuo Yang R T W 1978 Historical trends and statistics of Southern Oscillation, El Nino and Indonesian Droughts. *Fish. Bull.* 76: 663–678
- Rasmusson E M, Carpenter T H 1983 The relationship between eastern equatorial Pacific sea surface temperature and rainfall over India and Sri Lanka. *Mon. Weather Rev.* 111: 517–528
- Sikka D R 1980 Some aspects of large-scale fluctuations of summer monsoon rainfall over India in relation to fluctuations in the planetary and regional scale circulation parameters. *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* 89: 179–195

# Earthquake hazard in Indian region

K N KHATTRI

Wadia Institute of Himalayan Geology, Dehradun 248 001, India

**Abstract.** The causes of earthquakes and their incidence in the Indian region have been briefly reviewed. The basic elements of probabilistic seismic hazard analysis, which is a basis for making pragmatic decisions for strategies for hazard reduction and mitigation, are discussed. A probabilistic seismic hazard map of the Indian region, which delineates the seismic hazard in various regions in terms of peak ground accelerations expected to be exceeded with a probability of 10% in any 50-year period, is presented. The need and scope for further research are outlined.

**Keywords.** Earthquakes; seismic hazard analysis; seismic source zones; disaster mitigation strategies; faults; seismicity data; magnitude.

## 1. Introduction

India has suffered four great earthquakes of magnitudes 8.5 and greater, in the past hundred years, inflicting heavy casualties and economic damage. Yet, human memory being short, it is generally not recognized that we continue to live under the long shadow of such future calamities. One of the ways to mitigate the destructive impact of earthquakes is to conduct a seismic hazard analysis and take remedial measures.

Seismic hazard analysis consists of estimating various effects such as ground failure by faulting or soil liquefaction, ground shaking, tsunami generation etc. that might be caused by future earthquakes in a region.

The occurrence of earthquakes, for all practical purposes, is presently unpredictable with precision. Also, the effects of earthquakes such as ground shaking, ground failure etc. are quite variable. Thus, their analyses and parameterization have to be suitably done in a probabilistic framework. Strong ground shaking is the most widespread agent of earthquake hazard. Seismic hazard due to shaking is, therefore, the most frequently mapped, which provides probability estimates at given sites of the ground motion (displacement, velocity or acceleration) that will not be exceeded in the wake of future earthquakes in a given interval of time. Such parameterization can therefore form a rational basis for balanced decision-making in planning the whole gamut of hazard mitigation strategies.

Seismic hazards in a region may then also be translated into seismic risk by using a loss function that will include life as well as economic activity. In this paper, we review the construction of a probabilistic seismic hazard map of the country.

## 2. Causes of earthquakes

The immediate cause of a great majority of earthquakes is faulting. Stresses in the crust are steadily generated by the relative motion of plates, which are components of the fractured outermost shell of the earth. Cyclic release of these stresses owing to the finite strength of rocks and their reaccumulation provides a sustained environment for earthquake recurrence along plate margins, such as the Himalaya, as well as in the plate interior, such as the Indian shield, although the operative mechanisms there are not so well defined.

### Source zones

For seismic hazard analysis in a region, one begins with identifying the possible sources of future earthquakes. Past seismicity together with the model of regional tectonics is the basis for identifying seismic source zones. Since faults are the principal cause of earthquakes, one needs to identify such faults that are potential sources of future earthquakes. These faults have to be modelled in three dimensions. Furthermore, the random occurrence of smaller and moderate-sized earthquakes called background seismicity must also be considered in a source zone defined around active faults.

In many regions, earthquakes are caused by blind faults which have no direct expression on the surface and are difficult to model. In such places, one defines a source zone in which earthquakes can occur. The outlines of source zones are governed by aerial distribution of past earthquakes and geological considerations governing active tectonics. Figure 1 shows the seismicity of the Indian region and the identified seismic source zones (Khattri *et al* 1984).

## 3. Recurrence relation

The temporal distribution of earthquakes in a source zone is modelled by the recurrence relation,

$$\log_{10} N_c(M) = A - BM,$$

where  $N_c(M)$  is the number of earthquakes of magnitude equal to or greater than  $M$ , that occur during a unit time interval (say one year) in a unit area (say  $100 \text{ km}^2$ ),  $A$  and  $B$  being parameters specific to a particular source zone. This is the cumulative magnitude recurrence relation. The return period of an earthquake with magnitude  $M$  or greater is given by  $1/N(M)$  years.

Recurrence parameters are usually estimated by using a regression analysis of historical seismicity data. These are now being supplemented by paleoseismicity data, wherever possible, as they provide important constraints on the frequencies of large earthquakes that have long return periods. Historical seismicity data for such large and great earthquakes are often deficient owing to short time windows usually available, but a moving time-cum-magnitude window may be used to compensate for incompleteness of catalogs.

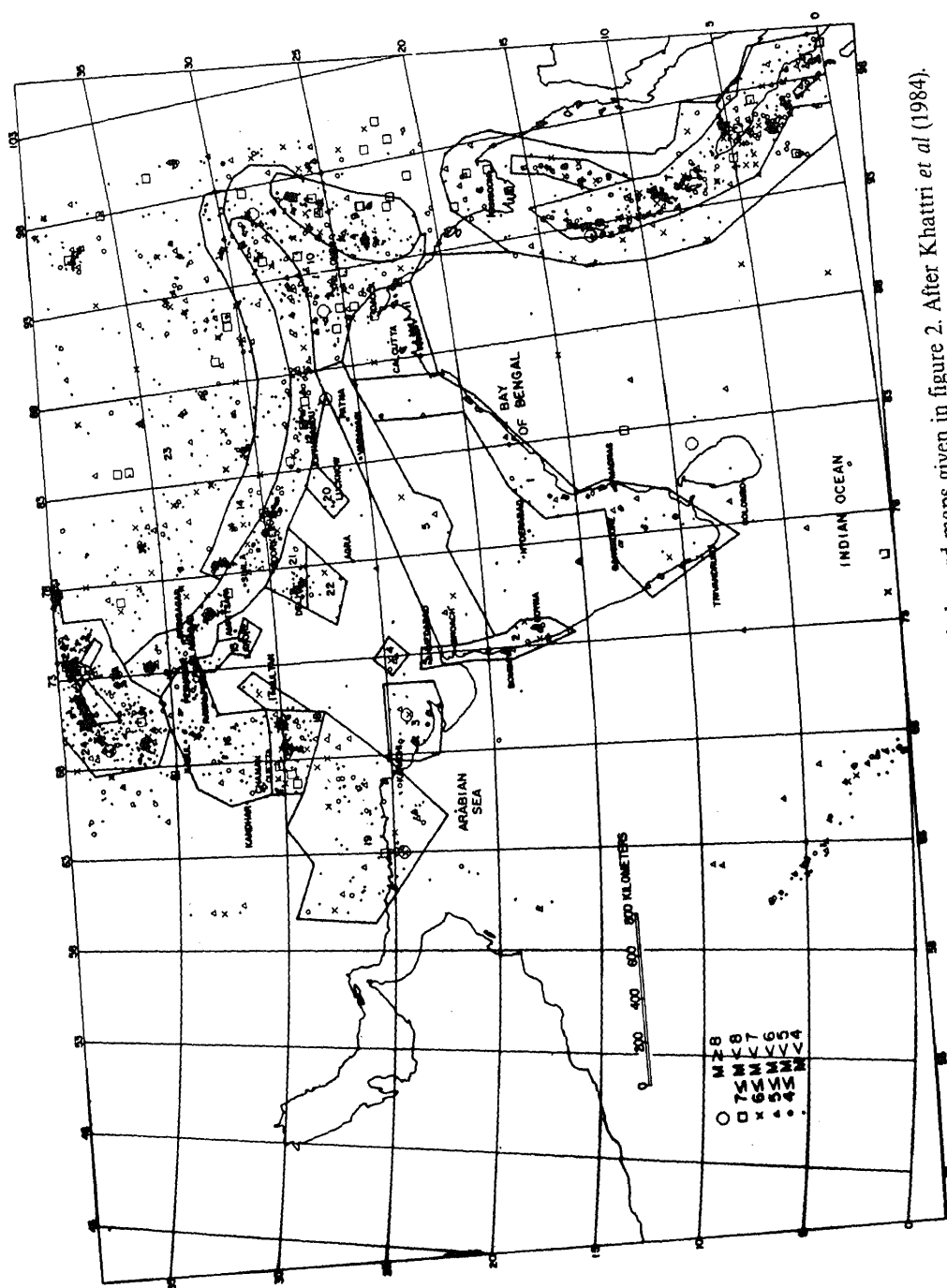
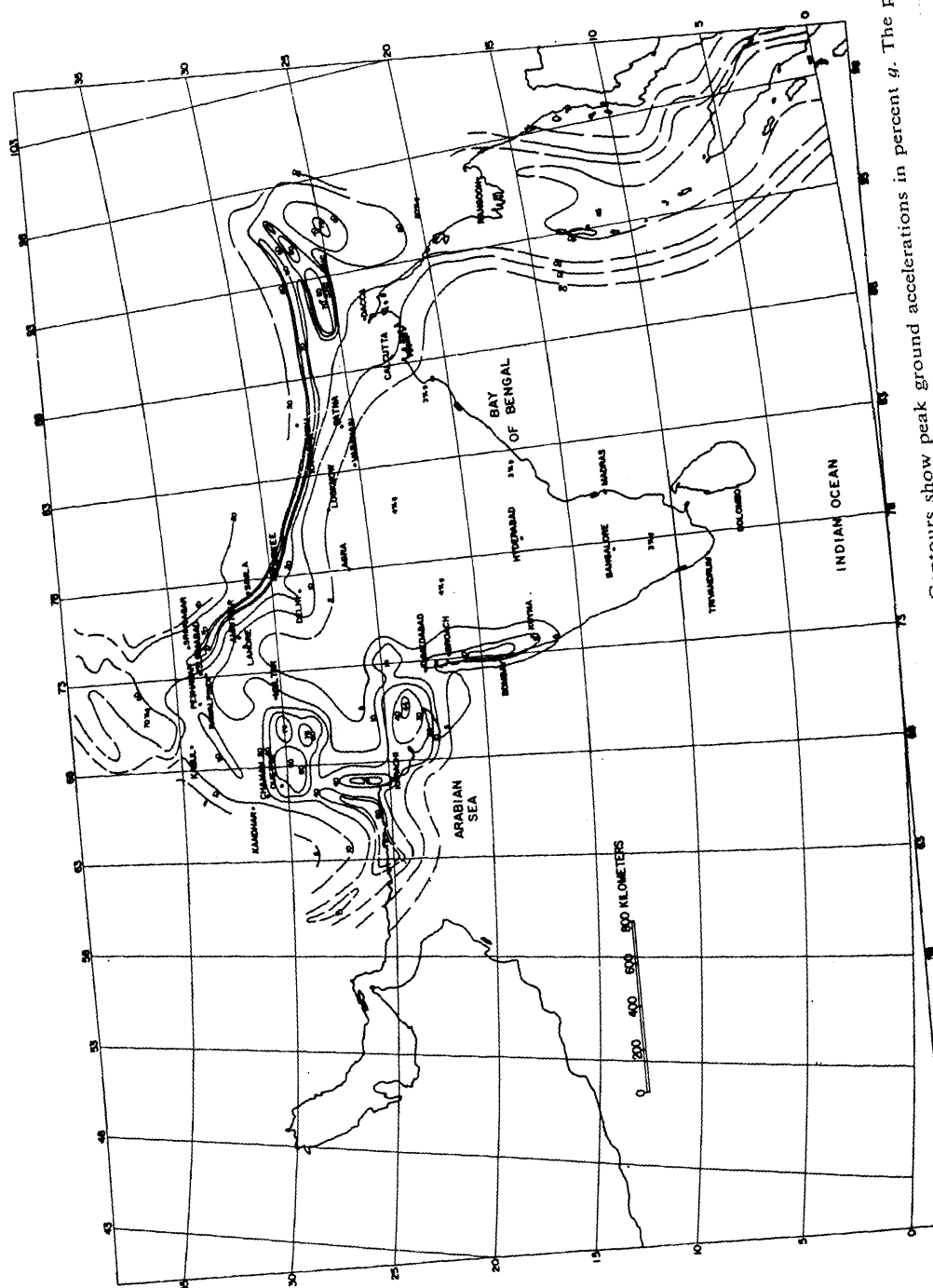


Figure 1. Seismic source zones used to develop seismic hazard maps given in figure 2. After Khattri et al (1984).



Contours show peak ground accelerations in percent  $g$ . The probability

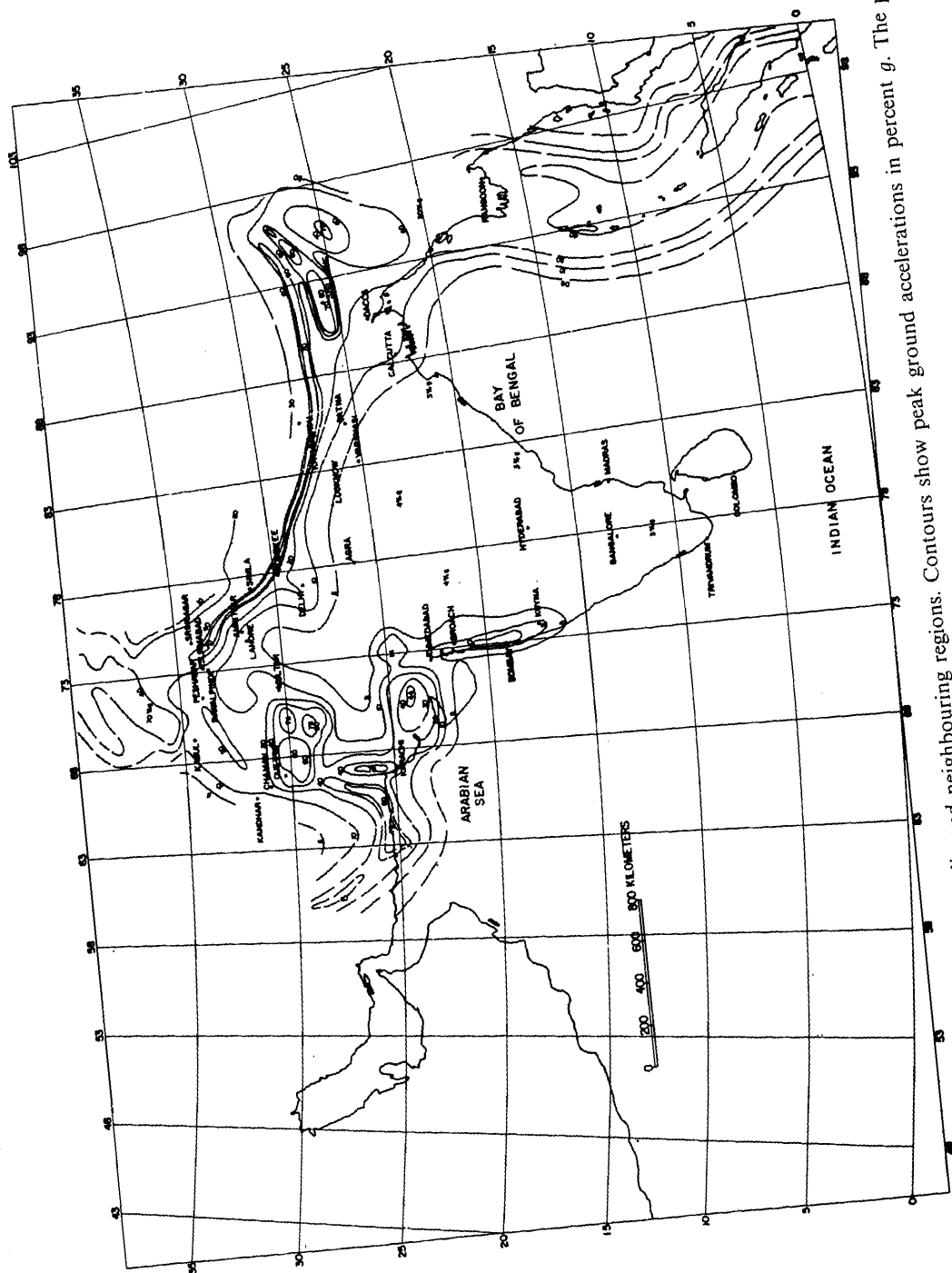


Figure 2. The seismic hazard map of India and neighbouring regions. Contours show peak ground accelerations in percent  $g$ . The probability of exceedence is 10% in 50 years. After Khattri *et al* (1984).

#### 4. Maximum magnitude earthquake

One also estimates the maximum possible earthquake magnitude for each zone. This is done on the basis of the model of tectonic processes responsible for seismicity in the region as well as the record of past earthquakes.

##### *Attenuation laws*

Since strong ground shaking is the immediate cause of widespread destruction during an earthquake, probabilistic hazard maps often map some single parameter to represent it. Usually, it is the peak ground acceleration or the peak ground velocity. The occurrence of an earthquake is thus translated into one of a peak ground motion parameter (usually acceleration) at various sites using empirical attenuation relationships of the type,

$$\log Y = a + b + d \log p[R + c(M)] + eR + G,$$

where  $Y$  is a given ground motion parameter,  $M$  the earthquake magnitude,  $R$  the source-to-site distance,  $c(M)$  is a term that reflects the behaviour of the relationship at small distances and large magnitudes, and  $G$  is a zero mean Gaussian random error term describing the variability in ground motion. The relation is established using regression analysis. It may be noted that prediction of  $Y$  is dependent on local and regional site geology.

#### 5. Modelling of extreme accelerations

Temporal sequence of earthquakes in a source region is usually modelled as a Poisson process. This model assumes the density distribution to be homogeneous in time, i.e. events are independent in time and are identically distributed. The distribution of earthquakes in space may be suitably modelled to reflect the nature of tectonics in the source region.

The extreme probability  $F_{\max,t}(\mathbf{a})$  of peak ground acceleration is mapped in probabilistic seismic hazard maps. This is the probability that at a particular site an observed peak ground acceleration  $\mathbf{a}$  will not be exceeded in a time window of  $t$  years. If earthquakes are represented by a Poisson distribution with a mean occurrence rate  $\lambda$ , the corresponding peak ground accelerations at a site will also follow the same distribution.

To introduce the time element, we let  $\lambda = pt$ , where  $p$  is the mean rate of occurrence of earthquakes per year and  $t$  is the time interval of interest. The extreme probability can then be written as (Khatti *et al* 1983),

$$F_{\max,t}(\mathbf{a}) = \exp\{-t/RP_Y(\mathbf{a})\}$$

where  $RP_Y$  is the return period in years of peak ground acceleration equal to or greater than  $\mathbf{a}$ .

The return period of maximum ground acceleration in an exposure period of 50 years, with a 90% probability of not being exceeded, is approximately 475 years.

A probabilistic earthquake hazard map of India and the surrounding region prepared in this way is shown in figure 2. The mapped peak ground accelerations

are in the neighbourhood of  $70\%g$  in the Himalaya which is the most vulnerable region of the country. Peak ground accelerations in Peninsular India by comparison are smaller, generally of the order of less than  $10\%g$ . Hazard is considered to be negligible for peak ground acceleration (pga)  $< 0.05g$ ; low for  $0.05g < \text{pga} < 0.1g$ ; moderate for  $0.1g < \text{pga} < 0.2g$ ; high for  $0.2g < \text{pga} < 0.4g$ ; and severe for  $0.4g < \text{pga}$ . We emphasize here that the map is for an exposure period of only 50 years. As the exposure period is increased the expected peak ground acceleration values will also increase. The return period of damaging earthquakes in the Himalaya is relatively small, of the order of a few years. In the peninsula, seismicity is feeble and damaging earthquakes are known to occur at considerably longer intervals. Their sizes are also relatively smaller. However, since the return periods of these earthquakes are long and unknown, there is considerable uncertainty about the accumulated strain levels in different parts of the peninsula, even as long periods of quiescence create an illusion that the region is safe from earthquakes. Meanwhile, centres of dense anthropogenic activities have become quite concentrated all over India, with hardly any preparation to face unsuspected earthquake hazards. This has been sadly brought home to us by the recent Uttarkashi and Killari (Maharashtra) earthquakes.

It is generally not appreciated that this country as a whole is not properly prepared to deal with earthquake hazards in a scientific and planned manner. Barely any systematic steps have been taken to prepare quantitative hazard maps on regional or local scales, not even for regions selected for the establishment of megacities or construction of large structures such as dams, nor has the use of concepts such as Bayesian estimates, compound probability distributions, information theoretic analysis, synthesis of earthquake-strong ground motion time histories, fractal descriptions, and newer data for seismic hazard maps been made to estimate hazards in the absence of adequate high quality data. Indeed, there is urgent need for a map depicting anticipated economic losses due to earthquakes and a plan for constructive social response in the wake of a damaging earthquake. Recent massive disasters caused by the moderate Uttarkashi and Killari earthquakes have underlined the grave necessity of designing and implementing a comprehensive earthquake disaster mitigation plan backed by hard research and sophisticated conceptual approaches.

## References

- Khattri K N, Rogers A M, Perkins D M 1983 Estimates of seismic hazard in northwestern India and neighbourhood. *Bull. Inst. Seismol. Earthquake Technol.* 20: 1–22
- Khattri K N, Rogers A M, Perkins D M, Algermissen S T 1984 A seismic hazard map of India and adjacent areas. *Tectonophysics* 108: 93–134



# Computational Heat Transfer

## Foreword

An important and active research area in engineering sciences is that of computational heat transfer. Considerable growth has occurred over the last three decades in the numerical techniques and methodology applicable to heat transfer as well as in the variety of thermal problems considered. A wide range of complexities and problems are encountered due to the application of thermal sciences to diverse engineering disciplines, ranging from environmental, energy, aerospace, and electronic systems to manufacturing and combustion processes. Consequently, there is a strong need to develop valid, accurate and versatile numerical schemes for simulating heat and mass transfer in complicated geometries, with realistic boundary conditions and with different material characteristics. Existing numerical methods have also been adapted to study different practical situations. Even though practical problems provide a major motivation for research in computational heat transfer, fundamental heat transfer processes such as convection, radiation, and conduction are also of interest since many basic questions need to be answered and a better understanding of the underlying phenomena is needed in many cases. Areas such as turbulent transport, conjugate heat transfer, effects of strong material property variations, three-dimensional processes and problems with phase change need further work.

This special issue of *Sādhanā* brings into focus some of the important aspects and techniques in computational heat transfer. The issue contains ten papers that cover many diverse and important topics. Most of the authors have also included a discussion of the current state of the art, besides presenting their own work, in order to provide the background for the topic and outline the current trends. Though the methodology is of particular interest in some of the papers, the application of numerical techniques is stressed in others. The subject area is a very vast one and it is not possible to consider all the different facets and techniques. However, this special issue of the journal strives to present some of the important research topics and the relevant numerical strategies which may be employed for obtaining accurate and valid solutions.

The first paper by Paolucci presents an excellent review on natural convection flow and heat transfer in an enclosed space, emphasizing important concerns such as non-Boussinesq effects, and transitional/turbulent flow. Current trends and future efforts needed on the numerical simulation of these flows are discussed. In the second paper, Mukutmoni and Yang further expand on the question of transitions in Rayleigh-Benard convection in rectangular enclosures, focusing on numerical studies concerning this complicated transport phenomenon. Temporal as well as spatial transitions are discussed, with detailed consideration given to pattern selection. In

the third paper, Lauriat and Desrayaud consider the computation of time-dependent laminar buoyancy-driven flows arising from horizontal wires in enclosed spaces. Transition to unsteady flows is simulated, considering different routes to chaos. Distinct regimes for the bifurcations are determined. Thus, these three papers are directed at the important questions of unsteadiness, bifurcations, transition and chaos in buoyancy-induced flow and transport. Finite difference and finite volume methods are considered by the first two papers, while the third one uses a finite element approach. The fourth paper by Molki and Faghri is also on natural convection, but on steady laminar flow arising from an interaction between the flow in an enclosure and an array of cavities. The complicated flow that arises and the resulting transport are examined using a finite volume scheme for a range of parameters, particularly the Prandtl number.

The fifth paper by Hoogendoorn *et al* discusses the numerical modelling of transport in furnaces for the thermal processing of glass. Turbulent flow and combustion in furnaces are included, implying modelling of radiative heat transfer, flame chemistry, soot formation and combustion products. Thus, the complexities due to species concentrations and combined modes is brought out for this practical circumstance. Validation of the model is based on comparisons with experimental data. The next paper, by Yücel *et al*, investigates combined natural convection and radiation in a square enclosure, using a finite volume scheme for the flow and the discrete ordinates method for radiation. The effect of radiation on the resulting temperature field and on the flow is shown to be substantial for a variety of circumstances. The heat fluxes at the walls are also computed.

A combined finite element and finite volume solution is discussed in the seventh paper, by Mohan and Tamma. Adaptive time stepping, based on the local and global error, is included to model transient problems in order to obtain optimal time steps during the entire analysis. A finite element solution of two- and three-dimensional internal flows is presented in the next paper by Srinivas *et al*. Turbulent flow modelling is also incorporated in the solution, considering nine different models for comparison. Flows in diffusers and ribbed channels are calculated by way of illustration. The next paper by Gopalakrishna and Jaluria investigates the complex transport phenomena associated with the thermal processing of chemically reactive polymers such as plastics and food materials. A finite difference scheme is developed to study the resulting heat and mass transfer in the channel of a single-screw extruder. The last paper by Date presents an enthalpy formulation to numerically model phase-change problems. This approach removes the problem of waviness in the temperature history which is often encountered in conventional formulations. A new method for tracking the interface is discussed.

It is evident that a range of important problems and techniques in computational heat transfer are presented in this special issue of the journal. The diversity of this important and interesting area of research is brought out, along with different types of numerical approaches that are commonly used, such as finite difference, finite element, spectral and control volume methods. Modifications in traditional schemes to meet the challenges posed by complexities that arise in practical problems are also brought out. The editors hope that this special issue will indicate the current trends and point out appropriate numerical methods for a variety of basic or applied problems in heat transfer.

We would like to thank the Editor, Professor N Viswanadham, for giving us the opportunity to organize this special issue. Computational heat transfer is an interesting and challenging area, which has seen tremendous growth in interest and activity in recent years. As such, it is particularly satisfying to see a special issue devoted to this topic.

October 1994

YOGESH JALURIA  
J SRINIVASAN  
Guest Editors



## The differentially heated cavity

S PAOLUCCI

Aerospace and Mechanical Engineering Department, University of Notre Dame, IN 46556–5637, USA

**Abstract.** This review discusses recent work dealing with natural convection flow in a differentially heated cavity. The emphasis is placed primarily on work dealing with the non-Boussinesq regime, transitional flow, and turbulent flow. Direction for future work in areas where additional effort is required is also provided.

**Keywords.** Natural convection; non-Boussinesq; transition; turbulence.

### 1. Introduction

Flow in rectangular enclosures are encountered in a variety of industrial applications. Because of their importance, such flows have been the subject of research for many decades. An attempt at a complete review and history of this subject is a formidable task which I will not attempt here. Instead, I will discuss the more recent results that have some technological importance and also reflect my current interest.

My discussion will deal exclusively with the study of natural convection flow within a vertical rectangular enclosure resulting from lateral heating in the presence of gravity. This is a prototypical problem that is relevant to many applications, such as thermal insulation of buildings, solar energy storage, crystal growth, and nuclear reactor core isolation. In many of these applications, direct modelling of the physical processes is rather complex. As a result, natural convection flow in this idealized configuration offers the opportunity for researchers to understand the more fundamental aspects of the resulting flows. In point of fact, this prototypical problem is considered as a benchmark in evaluating computational methods in the laminar (low Rayleigh number) regime. More importantly, this problem offers the opportunity to fully understand the transition mechanisms and gain substantial insight into natural convection turbulence. Since the flow is fully enclosed, boundary conditions are well defined. Ambient turbulent fluid fluctuations, that are difficult in general to characterize, do not enter in studies of flow transition and turbulence within the cavity. Furthermore, because the flow is fully bounded, no artificial (and generally incorrect) boundary conditions need be introduced in computational studies.

Excellent reviews of the earlier works on this problem are given by Ostrach (1972, pp. 161–227, 1982, pp. 365–79), Catton (1978, pp. 13–30), and Yang (1987, pp. 13.1–13.51, 1988). After discussing simplifying hypotheses that are often used in analyses of the problem, we will briefly review some of the earlier works, but more

emphasis is placed on more recent works dealing with non-Boussinesq effects, and studies that have shed some light on instabilities, transition, and turbulence within the cavity.

## 2. Problem definition

Consider a three-dimensional enclosure of width  $L$ , height  $H$ , and depth  $D$  containing a fluid as shown in figure 1. The fluid is initially quiescent and at a uniform temperature  $T_0^*$  and pressure  $p_0^*$ . The walls of the container are initially at the same temperature  $T_0^*$ . The  $x^*$ -coordinate is fixed on the left wall, and the  $y^*$ -coordinate is positive in the upwards direction. At times larger than zero the two vertical walls located at  $x^* = 0, L$  are maintained at temperatures  $T_h^*$  and  $T_c^*$  respectively, where  $\Delta T \equiv T_h^* - T_c^* > 0$ . Asterisk superscripts denote dimensional quantities.

We keep assumptions to a minimum at this stage and non-dimensionalize the problem with reference quantities for length, velocity, temperature, and thermodynamic pressure using the width  $L$ , the viscous diffusion speed  $u_r = \alpha_r/L$ , the average temperature  $T_r = (T_h^* + T_c^*)/2$ , and the initial pressure  $p_r = p_0^*$ , respectively, as follows:

$$\begin{aligned} x_i^* &= Lx_i, & t^* &= (L/u_r)t, & u_i^* &= u_ru_i, & \Pi^* &= \rho_ru_r^2\Pi, \\ \rho^* &= \rho_r\rho, & T^* &= T_rT, & \bar{p}^* &= p_r\bar{p}, & c_p^* &= c_{pr}c_p, \\ \beta^* &= \beta_r\beta, & \mu^* &= \mu_r\mu, & \lambda^* &= \mu_r\lambda, & k^* &= k_rk, \end{aligned} \quad (1)$$

where we have introduced a number of fluid properties, all of which are evaluated at the reference temperature and thermodynamic pressure.

The resulting dimensionless governing equations, valid under low Mach number conditions (generalized from those given by Paolucci 1982), but allowing for arbitrary density variations, are given as follows:

$$\partial\rho/\partial t + \partial\rho u_i/\partial x_i = 0, \quad (2a)$$

$$\frac{\partial\rho u_i}{\partial t} + \frac{\partial\rho u_j u_i}{\partial x_j} = -\frac{\partial\Pi}{\partial x_i} + \frac{\text{Ra Pr}}{2\varepsilon}\rho n_i + \text{Pr}\frac{\partial\tau_{ij}}{\partial x_j}, \quad (2b)$$

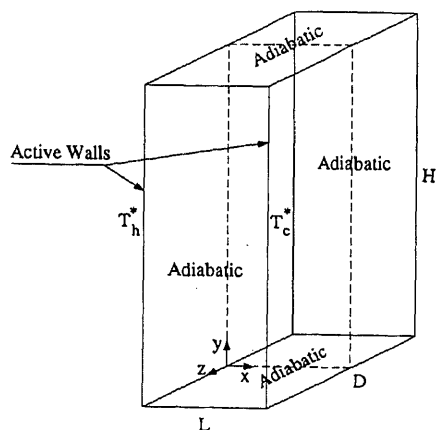


Figure 1. Schematic view of the cavity.

$$\rho c_p \left( \frac{\partial T}{\partial t} + u_j \frac{\partial T}{\partial x_j} \right) - \Gamma \beta T \frac{d\bar{p}}{dt} = \frac{\partial}{\partial x_j} \left( k \frac{\partial T}{\partial x_j} \right), \quad (2c)$$

$$\rho = \rho(\bar{p}, T), \quad (2d)$$

where  $u_i = (u, v, w)$  are velocity components in the  $x_i = (x, y, z)$  directions respectively,  $\Pi$  is the reduced pressure which accounts for the hydrostatic and hydrodynamic effect,  $n_i = (0, -1, 0)$  is the unit vector in the direction of gravity, and  $\tau_{ij}$  is the viscous stress tensor given by

$$\tau_{ij} = \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \lambda \delta_{ij} \frac{\partial u_k}{\partial x_k}, \quad (3)$$

where  $\lambda$  is the coefficient of bulk viscosity, and  $\delta_{ij}$  is the Kronecker delta.

The initial and boundary conditions expressed in dimensionless form are:

$$u_i = 0, \quad T = 1 \quad \text{at } t = 0,$$

$$u_i = 0, \quad T = 1 + \varepsilon \quad \text{at } x = 0 \quad \text{and} \quad u_i = 0, \quad T = 1 - \varepsilon \quad \text{at } x = 1,$$

$$u_i = 0, \quad \partial T / \partial y = 0 \quad \text{at } y = 0, A_H, \quad \text{and} \quad u_i = 0, \quad \partial T / \partial z = 0 \quad \text{at } z = \pm A_D / 2. \quad (4)$$

The spatially uniform pressure  $\bar{p} = \bar{p}(t)$  appearing in the energy equation and the equation of state accounts for the change of static pressure with time. The separation of pressure components is essential in removing acoustic waves from the equations, however this splitting introduces  $\bar{p}$  as an extra unknown. Now the general equation of state can be rewritten more explicitly as

$$\rho = \exp \left( - \int_1^T \beta dT' + \int_1^{\bar{p}} \kappa d\bar{p}' \right), \quad (5)$$

where  $\beta = -(\partial \rho / \partial T)_{\bar{p}} / \rho$  is the coefficient of volume expansion, and  $\kappa = (\partial \rho / \partial \bar{p})_T / \rho$  is the isothermal compressibility coefficient. Using the global mass conservation statement

$$\frac{d}{dt} \int_V \rho dV = 0, \quad (6)$$

in conjunction with local continuity and energy equations, we obtain the following differential-integral equation for the static pressure,

$$\begin{aligned} \frac{d\bar{p}}{dt} = & \left\{ \int_S k \frac{\partial T}{\partial x_j} dS_j + \int_V u_i \frac{\partial}{\partial x_i} \left[ \rho c_p / \left( \beta - \int_1^{\bar{p}} \frac{\partial \kappa}{\partial T} d\bar{p}' \right) \right] dV \right\} \div \\ & \left\{ \int_V \left[ \rho c_p \left( \kappa - \int_1^T \frac{\partial \beta}{\partial \bar{p}} dT' \right) / \left( \beta - \int_1^{\bar{p}} \frac{\partial \kappa}{\partial T} d\bar{p}' \right) - \Gamma \right] dV \right\}, \quad (7) \end{aligned}$$

where  $S$  and  $V$  are the surface area and volume of the cavity. This equation is complemented by the initial condition

$$\bar{p} = 1 \quad \text{at } t = 0. \quad (8)$$

The independent dimensionless parameters appearing in the problem are:

$$\text{Ra} = \beta_r g \Delta T L^3 / \nu_r \alpha_r, \quad \text{Pr} = \nu_r / \alpha_r, \quad \varepsilon = \frac{1}{2} \beta_r \Delta T, \\ \Gamma = (1/\sigma_r T_r)((\gamma_r - 1)/\gamma_r), \quad A_H = H/L, \quad A_D = D/L. \quad (9)$$

In the above definitions and non-dimensionalizations  $\rho$  is the density;  $\alpha = k/\rho c_p$  and  $\nu = \mu/\rho$  are the kinematic viscosity and thermal diffusivity, while  $\mu$  and  $k$  are the dynamic viscosity and thermal conductivity;  $c_p$ ,  $c_v$  and  $\gamma$  are the specific heats at constant pressure and volume and their ratio;  $\sigma = (\partial \bar{p}/\partial T)_\rho/\bar{p}$  is the coefficient of tension; and lastly  $g$  is the magnitude of the gravitational field.

We emphasize the fact that (2), (3), and (7) are applicable to the natural convection flow of any fluid within a fully enclosed cavity. No assumptions regarding property variations are made. Solutions of these equations account for all non-Boussinesq effects in natural convection. Most non-Boussinesq results to date are for air, whose flow may accurately be described by using the Stokes assumption  $\lambda = -\frac{2}{3}\mu$  and the thermodynamic properties by the calorically perfect gas assumptions  $c_p = 1$ ,  $\beta = 1/T$ ,  $\kappa = 1/\bar{p}$ , and  $\sigma = 1/T$ . Subsequently, we have  $\Gamma = (\gamma_r - 1)/\gamma_r$ , and (5) and (7) simplify to

$$\rho = \bar{p}/T, \quad (10)$$

$$\frac{d\bar{p}}{dt} = \frac{\gamma_r}{V} \int_S k \frac{\partial T}{\partial x_j} dS_j. \quad (11)$$

In addition, for air  $\gamma_r = 7/5$ ,  $\text{Pr} = 0.71$ , and the transport properties are accurately approximated by the Sutherland law models

$$\mu = T^{3/2} \left( \frac{1 + S_\mu}{T + S_\mu} \right), \quad k = T^{3/2} \left( \frac{1 + S_k}{T + S_k} \right), \quad (12)$$

where using  $T_r = 300$  K and normal pressure, the dimensionless Sutherland constants are  $S_\mu = S_\mu^*/T_r = 0.368$  and  $S_k = S_k^*/T_r = 0.648$  (see White 1974). From the definition of  $\varepsilon$  we note that  $0 < \varepsilon < 1$  corresponds to the temperature difference range of  $0 < \Delta T < \infty$ . Obviously the range of validity of the Sutherland law is considerably less. As a point of reference, note that  $\varepsilon = 0.6$  corresponds to  $T_h^* = 480$  K and  $T_c^* = 120$  K for  $T_r = 300$  K, and represents a practical upper limit on the validity of the results for air resulting from increasing errors in the Sutherland law conductivity at the cold wall (see Chenoweth & Paolucci 1986).

It was shown by Paolucci (1982) that in the limit  $\varepsilon \ll 1$ , we have  $\bar{p} = 1$ ,  $d\bar{p}/dt = 0$ , and subtracting out the hydrostatic components, (2), (3), and (5) reduce to the classical Boussinesq equations

$$\frac{\partial u_i}{\partial x_i} = 0, \quad (13a)$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{\partial \Pi_d}{\partial x_i} - \text{RaPr} \theta n_i + \text{Pr} \frac{\partial^2 u_i}{\partial x_j \partial x_j}, \quad (13b)$$

$$\frac{\partial \theta}{\partial t} + u_j \frac{\partial \theta}{\partial x_j} = \frac{\partial^2 \theta}{\partial x_j \partial x_j}, \quad (13c)$$

where  $\Pi_d$  is the dynamic component of pressure and  $\theta = (T - 1)/2E$ . It is emphasized



that the Boussinesq equations will only yield relatively accurate solutions only for small temperature differences. For example, Gray & Giorgini (1976) show that the maximum temperature differences for air and water for which these equations are applicable are  $28.6^\circ$  and  $1.25^\circ\text{C}$  respectively.

Solutions for the Boussinesq equations have been obtained for a large range of Prandtl numbers and primarily in a two-dimensional cavity where  $A_D \rightarrow \infty$ . Furthermore, some solutions have been obtained using perfectly conducting top and bottom boundaries so that the thermal conditions at  $y = 0$ ,  $A_H$  in (4) are replaced by

$$T = 1 + \varepsilon(1 - 2x). \quad (14)$$

The problem studied here is perhaps the simplest one possible involving convection heat transfer between two isothermal boundaries; in spite of this apparent simplicity, the physical nature of the resulting flow regions, even at steady-state, can be an incredibly complex function of the controlling independent dimensionless parameters.

### 3. Numerical techniques

The Boussinesq equations have been solved in both streamfunction-vorticity and primitive variable formulations. Since most of the recent results deal with non-Boussinesq effects, and since any computational scheme that one hopes to apply to transition and turbulence studies must be easily extendable to three-dimensions, most of the results that will be discussed have been obtained using a primitive variable formulation.

The most accurate simulations to date have used a primitive variable formulation in conjunction with either a pseudo-spectral or a control-volume spatial approximation combined with a finite-difference time integration.

The principle of pseudo-spectral methods is the expansion of a variable in a finite series of orthogonal trial solutions. For the present problem where we have rigid walls one may use Chebyshev polynomials. In the pseudo-spectral method the Chebyshev polynomials are evaluated at specific spatial locations called the Gauss-Lobatto points. The choice of these collocation points simplifies the computation of the nonlinear terms and allows the use of Fast Fourier Transforms. The main advantage of pseudo-spectral methods is their accurate approximation of sufficiently smooth functions. However, they are restricted to simple geometries, the grid refinement distribution in the boundary layers is only trigonometric, and the construction of an efficient code is quite complicated.

Therefore, finite-difference or finite-volume methods are preferred by most researchers. However, their application must meet strict requirements. The scheme must be able to resolve accurately all scales of motion down to the size of a few grid cells. Moreover, there must be none or little numerical diffusion. These requirements rule out standard upwind schemes, which are first order and have strong numerical diffusion. Therefore, numerical schemes with at least second-order accuracy are mandatory. These schemes are usually applied on a staggered grid located on or within a finite control volume. The velocity components are defined at the centre of the sides whose normals point in the corresponding component directions, while scalar quantities are defined at the centre of the volume. For the finite difference formulation of the nonlinear convective terms one should adopt an energy conserving

scheme to avoid nonlinear instability (see Arakawa 1966, and Piacsek & Williams 1970). The time advancement is frequently calculated by either the Euler or the Adams–Bashforth method.

To enforce the condition of incompressibility in the Boussinesq limit, the continuity equation is reformulated in the form of a Poisson equation for the dynamic pressure (see Peyret & Taylor 1983),

$$\frac{\partial^2 \Pi}{\partial x_i \partial x_i} = \frac{\partial R_i}{\partial x_i} - \frac{\partial}{\partial t} \left( \frac{\partial u_i}{\partial x_i} \right), \quad (15)$$

in which  $R_i$  contains the contributions of the nonlinear terms. To calculate the pressure at time step  $n$  the last term on the right-hand side of (15) is approximated by

$$\frac{\partial}{\partial t} \left( \frac{\partial u_i}{\partial x_i} \right) \approx \frac{1}{\Delta t} \left[ \left( \frac{\partial u_i}{\partial x_i} \right)^{n+1} - \left( \frac{\partial u_i}{\partial x_i} \right)^n \right].$$

The first term on the right-hand side is set equal to zero in agreement with continuity. The second term is kept as a correction in order to force the solution to a divergence-free flow field (see Deardorff 1973, pp. 271–311).

For details of analogous algorithms applied to the non-Boussinesq equations the reader is referred to Chenoweth & Paolucci (1986) for the control volume formulation and to Le Quéré *et al* (1992) for the pseudo-spectral formulation.

Next we consider the boundary conditions. On the fixed walls one must use the no-slip condition for the velocity and the appropriate temperature conditions. In solving the discretized form of the Poisson equation, knowledge of the pressure gradient normal to the walls is required. A simple and popular choice is to set this gradient to zero similar to boundary layer flows. For physical reasons no boundary conditions for the pressure ought to be prescribed at the walls. A zero gradient normal to the walls, however, is not a bad approximation as long as the Rayleigh number is high so that boundary layers exist, and as long as the boundary layers are neither separated nor leaving the surface. Such is not the case in all wall regions of the cavity. A consistent and correct solution for the pressure gradient is obtained by evaluating the momentum equations at the walls.

In a direct simulation of the transition or turbulent flow, the flow field must remain completely resolved up to the wall. This sets important restrictions on the grid spacing. Therefore, one must resort to a non-uniform grid near the walls.

To conclude this section we briefly discuss the initial conditions. Most simulations of laminar convection are started from the quiescent isothermal conditions expressed in (4) (and (8) in solving the non-Boussinesq equations). However, to expedite calculations, a different approach is used in transition studies. In this case, the initial conditions of an integration for a certain value of the Rayleigh number is obtained from the solution corresponding to a Rayleigh number which is close to it.

#### 4. Cavities with $A_H \geq 1$

Although gas gaps between vertical parallel walls have been used for many decades to reduce heat transfer, their use with large horizontal temperature differences has become increasingly important during the last two decades. Examples of typical applications include insulation using double-pane windows or double walls, nuclear

reactors, fire in buildings, solar collectors, and electronic components in enclosures. Still the majority of the published investigations involve small temperature differences where gas properties are taken to be constant and the Boussinesq limiting equations are valid. Here we briefly describe some of the known non-Boussinesq analytical, experimental, and numerical results in the laminar regime.

The first model of the flow in the cavity was proposed by Batchelor (1954) for small values of Rayleigh number. This model is not very valuable since these small values are well below values of practical significance. Gill (1966) subsequently proposed a boundary layer model in a two-dimensional cavity having large aspect ratios ( $A_H \gg 1$ ) and containing a fluid of large Prandtl number. Bejan (1979) completed Gill's boundary layer model by imposing adiabatic boundary conditions (and zero global heat flux) on the top and bottom boundaries, even though the boundary layer is not valid on these horizontal walls. The experimental studies of Eckert & Carlson (1961) using air and Elder (1965) using silicone oil ( $Pr \approx 10^3$ ) contributed a great deal to understanding the flow in cavities whose vertical aspect ratios varied between 2.1 and 46.7 and the Rayleigh numbers between  $2 \times 10^2$  and  $2 \times 10^5$ . The work of Gill (1966) modified by Bejan (1979), was extended by Graebel (1981) who examined the influence of the Prandtl number. He demonstrated that the flow is only weakly influenced when the Prandtl number is of order unity.

Hara (1958), and Sparrow & Gregg (1958) gave analytical results for a vertical flat plate with large temperature differences. In general their results are not applicable to the vertical slot; however, Chenoweth & Paolucci (1986) have shown how their heat transfer results can be adapted to the differentially heated cavity. Polezhaev (1967) used equal Sutherland laws for dynamic viscosity  $\mu$  and thermal conductivity  $k$  (constant Prandtl number when the heat capacity at constant pressure is constant) in his numerical solution of the compressible Navier-Stokes equations; he used a relatively small temperature difference parameter of  $\varepsilon = 0.2$  and slot aspect ratios of unity. Rubel & Landis (1970, pp. 1-11) made an expansion in terms of  $\varepsilon$  to obtain first-order corrections to the zeroth-order Boussinesq results. They included fluid property variation by means of temperature power-law expressions and they assumed that the pressure was independent of the temperature difference as well as other parameters. Leonardi & Reizes (1979, pp. 297-306, 1981, pp. 387-412) also numerically solved the compressible Navier-Stokes equations using equal Sutherland laws for  $\mu$  and  $k$ ; however they did examine cases where larger temperature differences were involved ( $\varepsilon \leq 0.6$ ), and they studied only aspect ratios of 1 and 2, although some of their results are suspect (see Chenoweth & Paolucci 1986).

Experimental studies using gases with large temperature differences have been equally sparse. Eckert & Carlson (1961) did investigate the high aspect ratio problem, but they only examined cases where  $\varepsilon \leq 0.13$ , so that the Boussinesq solution is approximately valid for their results. Similarly, all of the Mordchelles-Regnier & Kaplan (1963, pp. 94-111) experiments in the laminar region were in the Boussinesq regime, although they did obtain some turbulent non-Boussinesq results for a single flat plate. More recently, Duxbury (1979) obtained experimental results covering a wide parameter range in the steady laminar regime, but with  $\varepsilon \leq 0.15$ . It is not clear to what extent the above experimental studies were affected by heat losses from the ends.

Chenoweth & Paolucci (1985) investigated the steady-state fully developed boundary-layer region for large temperature differences between vertical isothermal walls. They derived some exact laminar solutions to the Navier-Stokes equations

for perfect gases with the properties described by unequal Sutherland laws. They produced variable Prandtl number results for  $0 < \varepsilon < 1$  using air as an example, although the accuracy of the Sutherland law conductivity degrades rapidly above  $\varepsilon = 0.6$ .

Varying amounts of theoretical and experimental information relating to flow region classification in the Boussinesq limit and near Boussinesq limit can be found in Batchelor (1954), Bergholz (1978), Duxbury (1979), Eckert & Carlson (1961), Lauriat (1980), Lee & Korpela (1983), Mordchelles-Regnier & Kaplan (1963, pp. 94–111), Polezhaev (1968), and Yin *et al* (1978). Much of the difference in the flow region bounds often can be related to the use of different criteria for the classification, although in some cases poor numerical resolution or experimental difficulties are responsible for the differences. A somewhat different picture will emerge if velocity field rather than thermal field or heat transfer information is used. Chenoweth & Paolucci (1986) present results for air using the ideal gas law and Sutherland law transport properties. Numerical solutions of the transient Navier–Stokes equations are used to generate laminar steady-state results primarily in the independent boundary layer region and the developing merged boundary layer region. However, other flow regions are also covered to the extent necessary to construct a better understanding of the entire laminar parameter range for all aspect ratios greater than or equal to unity. The velocity field behaviour is used to classify the different flow regions. The need to construct a map of flow regions in parameter space, which includes stationary and oscillatory stability boundaries, necessitated the use of the transient form of the equations. Wide ranges of aspect ratio, Rayleigh number, and temperature-difference parameters are examined. The results are compared in detail to the exact solution in the conduction and fully developed merged boundary layer limits for arbitrary temperature difference, and to the well-established Boussinesq limit for small temperature difference.

Figure 2 presents the flow region maps for  $Pr = 0.71$  obtained from their extensive computations. The solid lines for the Boussinesq limit  $\varepsilon \ll 1$  and the dashed lines for  $\varepsilon = 0.6$ , which bound some of the regions, denote lines of stability. The limits of the boundary layer regimes are given by the dot-dash and the dotted lines, for  $\varepsilon \ll 1$  and  $\varepsilon = 0.6$ , respectively. The shaded region represents the unsteady transition to

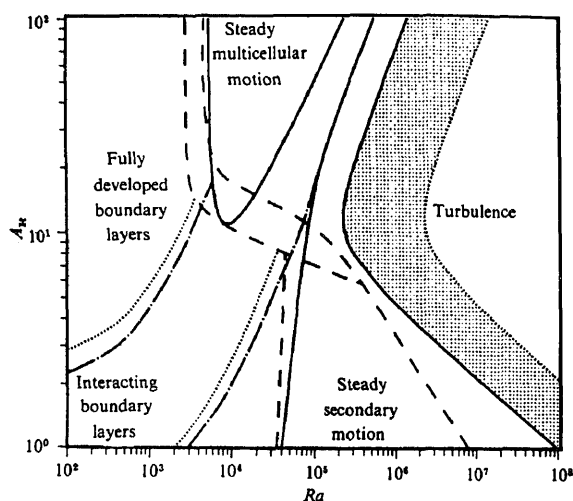


Figure 2. Flow regions dependence on  $A_H$ ,  $Ra$  and  $\varepsilon$  for  $Pr = 0.71$ .

turbulence for  $\varepsilon \ll 1$ . Correlations defining boundaries of the regions shown in figure 2 are given by Chenoweth & Paolucci (1986) and are discussed below.

The most familiar flow region, often called the boundary-layer region, is characterized by independent (non-interacting) boundary layers on the hot and cold walls; here the boundary layers are separated by a nearly stagnant core which always displays an approximately linear vertical thermal stratification. The boundary layer region exists primarily (although not exclusively) for low aspect ratios,

$$1 \leq A_H \leq 12(1 + 4\varepsilon)/Ra^{0.25\varepsilon}, \quad (17)$$

and for Rayleigh numbers in the range

$$7 \times 10^3(A_H - 0.6) < Ra < 10^4 A_H \left( 1.5 + \frac{10^4}{A_H^4} \right), \quad (18)$$

for  $\varepsilon \ll 1$ , and

$$4.5 \times 10^3(A_H - 0.6) < Ra < \frac{8 \times 10^6 A_H^{-1.7} + 5 \times 10^3(A_H/12)^6}{1 + (A_H/12)^6}, \quad (19)$$

for  $\varepsilon = 0.6$ . The upper limits in (18) and (19) represent critical Rayleigh numbers, since above those limits the flow no longer reaches a steady-state. When  $Ra$  is increased significantly above these critical values, the flow becomes increasingly unsteady and eventually becomes turbulent.

Below the lower limit of the independent boundary-layer region and for

$$400[A_H(1 - \varepsilon^3) - 2] < Ra < 7 \times 10^3(1 - \varepsilon^2)(A_H - 0.6), \quad (20)$$

and aspect ratios approximately in the range (17), the velocity boundary layers are interacting to an increasing degree as the value of  $Ra$  is decreased. Finally for

$$Ra < 400[A_H(1 - \varepsilon^3) - 2], \quad (21)$$

fully developed merged boundary layers exist near the mid-height of the slot. At that point both velocity and temperature are within 3% of universal profiles which are independent of the vertical coordinate. As the aspect ratio is increased further, the fully developed profiles exist in a region whose size is proportional to  $A_H$ , since the development regions near each end remain nearly fixed in size for given  $Ra$  and  $\varepsilon$ . The fully developed region is centred about  $A_H/2$  for small  $Ra$ , but for non-Boussinesq cases, when  $Ra$  is near the critical value of  $5760(1 + 0.434\varepsilon)^{-3}$  and  $A_H > 11$ , it may be centred significantly below  $A_H/2$ , since near that limit the bottom development region is approximately half that found at the top of the slot. For aspect ratios greater than 12–15, if the critical Rayleigh number of  $5760(1 + 0.434\varepsilon)^{-3}$  is exceeded, another type of instability can appear. This instability is characterized by standing waves and gives rise to multicellular flow inside the primary roll. Then combinations of large and small cells cover the entire slot and increase in strength with increasing  $Ra$ . The number and size of these cells depend on  $A_H$ ,  $Ra$  and  $\varepsilon$ . We note that in the Boussinesq limit these cells are approximately the same size, their number depends only on the aspect ratio, and their strength increases with increasing  $Ra$  (Lee & Korpela 1983). This multicellular motion is completely steady.

In the Boussinesq limit, as the Rayleigh number is further increased, a region of

reverse transition from multicellular to unicellular flow exists. This region is bounded by the multicellular region described by

$$A_H > 4.2 \times 10^{-3} Ra^{1.31} (Ra - 5760)^{-1/2}, \quad (22)$$

and by a thin boundary layer region containing steady secondary motion described by

$$4 \times 10^4 A_H^{1/4} (1 + 10^{-2} A_H^{5/4}) < Ra < 10^4 A_H (1.5 + (10^4/A_H^4)). \quad (23)$$

Above the upper limit of (23), unsteady motion exists. The region of secondary motion described by (23) is in fact also valid for relatively low aspect ratios as shown in figure 2 and occupies most of the boundary layer region defined by (18). This is also true for  $\varepsilon = 0.6$  since for  $A_H < 8$  the lower limit of (23) is then replaced by

$$Ra > 3.5 \times 10^4 A_H^{1/8}. \quad (24)$$

The reverse transition region, whose boundaries merge as  $A_H \rightarrow \infty$  for  $\varepsilon \ll 1$ , does not exist for  $\varepsilon = 0.6$  since in this case the boundaries merge near the upper limit of (17) as shown in figure 2. In fact, in this case, the approximate expression (22) and the lower limit of (23) are now replaced by

$$Ra > 2880 + (40/A_H)^{6.67}, \quad (25)$$

so that strong multicellular motion is now present in this entire region. This large change for  $A_H \gg 1$  and for  $\varepsilon = 0.6$  is accompanied by a greatly reduced critical Rayleigh number for unsteady motion as indicated by the upper limit of (19). The steady multicellular region is now confined to a narrow region between (25) and the upper limit of (19).

The location of the unsteady transition, displayed in figure 2, is defined by the upper limit of (18). The results of Mordchelles-Regnier & Kaplan (1963, pp. 94–111) for  $A_H < 12$  and extending down to aspect ratios as low as 1.67, predict that the critical Rayleigh number varies with  $A_H^{-1/2}$ . Most of the disagreement with the results of Chenoweth & Paolucci (1986) is at the lower aspect ratios where they show an  $A_H^{-3}$  dependence. We note that as the aspect ratio is decreased, heat losses from the ends become more and more important and cause instabilities to occur at a lower value of Rayleigh number. Apparently, these end losses are responsible for the Mordchelles-Regnier & Kaplan (1963, pp. 94–111) weaker dependence on aspect ratio. Quantitative agreement is much better with results obtained by linear stability analysis by Bergholz (1978). In making this comparison, use is made of the fact that the temperature in the core is approximately linearly stratified with a stratification parameter which is near unity in this region. Bergholz's results do show that the critical Rayleigh number varies approximately with aspect ratio as  $A_H^{-3}$  for  $A_H < 6$ . The result obtained from Bergholz's analysis arises from a travelling wave instability along the side walls. However, Chenoweth & Paolucci (1986) offer another possible explanation for the  $A_H^{-3}$  behaviour due to a "hydraulic" jump instability on the end walls. The experimental results of Ivey (1984) for  $Pr = 7.1$  and  $A_H = 1$  clearly indicate the presence of a "hydraulic" jump and a stationary wavetrain downstream of it when the Rayleigh number is above the critical internal Froude number. His observations are in complete agreement with those of Paolucci & Chenoweth (1989).

Bergholz's (1978) results also show that there is a minimum critical Rayleigh number for oscillatory instability, occurring at  $1.15 \times 10^5$  at an aspect ratio near 12.

This compares well with the minimum of  $Ra = 2.4 \times 10^5$  also occurring near  $A_H = 12$  shown in figure 2. At aspect ratios larger than 12 the results of Chenoweth & Paolucci (1986) show that the stratification parameter is nearer 0.5, and the critical Rayleigh number is approximately linear with  $A_H$ . This picture is in good agreement with the travelling wave instability from steady to unsteady motion given by Bergholz (1978), and consistent with experimental data of Mordchelles-Regnier & Kaplan (1963, pp. 94–111) for  $12 \leq A_H \leq 25$ .

Results of Lee & Korpela (1983), and Lauriat (1980) show that in the Boussinesq limit, for  $Ra \approx 10^4$  the flow changes from unicellular to multicellular motion as the aspect ratio is increased. They observe the transition occurring near  $A_H = 11 - 12$ , in agreement with the minimum aspect ratio shown in figure 2.

Still in the Boussinesq regime, Bergholz (1978) also obtained the critical Rayleigh number corresponding to the standing wave or stationary transition from fully developed merged boundary layers to steady multicellular motion. He obtained this value as approximately 5740, which again compares very well with the results,

$$Ra < Ra_c = 5760(1 + 0.434\epsilon)^{-3}, \quad (26)$$

for

$$A_H > A_{Ha} = (2 + Ra/400)/(1 - \epsilon^3), \quad (27)$$

given by Chenoweth & Paolucci (1986) in the limit  $\epsilon \rightarrow 0$ . However, Lee & Korpela (1983), using 17 grid points in the horizontal direction, obtain this transition between 7100 and 7810 for  $A_H = 20$ , which is much higher than the result shown in the figure and Bergholz's result. They explain this disagreement by noting that for large but finite aspect ratios of order 20, there is a small positive vertical stratification through the cavity, since the boundary layers are not yet completely developed, and Bergholz shows that the flow is stabilized with increasing stratification. Even though this argument correctly explains why the critical Rayleigh should increase for lower aspect ratios, figure 2 does not show a substantial increase until  $A_H < 20$ . In fact, the calculations of Chenoweth & Paolucci (1986) for  $Ra = 7100$  and  $A_H = 20$  show multicellular motion, in agreement with the results of Roux *et al* (1980) obtained by using a higher order scheme. A better explanation for the disagreement is that Lee & Korpela (1983) lacked sufficient horizontal resolution. The lower bound on the fully developed boundary layer region  $A_H > Ra/500$  given by Batchelor (1954) agrees with figure 2 for large  $Ra$  but differs substantially at low  $Ra$  since fully developed boundary layers at the mid-height plane cannot be developed for  $A_H < 2$  even if  $Ra$  and  $\epsilon$  approach zero.

Polezhaev (1968) constructed a somewhat complete picture of the flow regions for the same  $A_H$  and  $Ra$  ranges as shown on figure 2. In spite of the very coarse grids used, the regions where calculations were made and those where he used data from Eckert & Carlson (1961) as well as Elder (1965a) are in qualitative agreement with figure 2. However, in the regions where he extrapolated the bounds, there is substantial disagreement, particularly where both  $A_H$  and  $Ra$  are large and where both are small. The same statement applies to the regions plot given by Yin *et al* (1978), since only a band of results extending from low  $Ra$  and high  $A_H$  to high  $Ra$  and low  $A_H$  were used to locate lines which were then extrapolated across the other regions.

Roux *et al* (1980) found, in the Boussinesq limit, that a small region of reverse transition from multicellular to unicellular flow exists within the lower right corner of the region defined by  $A_H \geq 12$  and  $10^4 \leq Ra \leq 10^5$ . We note that this region

corresponds to the narrow region separating stationary and oscillatory stability branches, for a narrow range of stratification parameter, as given in figure 3 of Bergholz (1978). Chenoweth & Paolucci (1986) have verified the existence of this narrow region for  $\varepsilon \ll 1$ . However, as  $\varepsilon$  increases, this region quickly disappears, and as suggested by the dashed line in figure 2, it does not exist for  $\varepsilon = 0.6$ .

The influence of the Prandtl number is found to be minor when  $Pr > 1$ . For  $Pr < 1$ , however, the velocity field and heat transfer rates have a strong dependence on  $Pr$ . Quon (1972) performed finite difference calculations for the boundary layer regime in a square cavity in the Boussinesq limit. Consistent with the results of MacGregor & Emery (1969) he showed that the flow in this regime is insensitive to the magnitude of the Prandtl number for  $7.14 < Pr < 900$ . Korpela *et al* (1973) report that when the Rayleigh number reaches the critical value, the conduction regime may become unstable in two ways. For  $Pr < 12.7$ , the instability sets in as vertical stationary cells, with the critical Rayleigh number nearly independent of  $Pr$ . When  $Pr > 12.7$ , the instability is manifested in the form of travelling waves. More recently, Lee & Korpela (1983) carried out extensive numerical computations of multicellular flows in cavities of aspect ratios as large as 40 for a wide range of Prandtl numbers in the Boussinesq limit. They found that for low Prandtl number fluids ( $Pr \ll 1$ ) such as liquid metals, multicellular flows can develop when the aspect ratio is as low as 6. For air ( $Pr = 0.71$ ), the aspect ratio must be at least 12 before the flow becomes more complex as a result of instability. These results are consistent with the stability results of Bergholz (1978).

Very few numerical calculations have been performed in three-dimensional cavities. Of these we note the works of Mallinson & de Vahl Davis (1977) and Lee *et al* (1988) with cavities having  $A_H = 1$  and  $A_D$  varying from 2 to 4, and the most recent results of Fusegi *et al* (1991) with  $A_H = A_D = 1$ . Because of the large computational costs, all such calculations have been limited to  $Ra \leq 10^6$ . The exception is the work of Lankhorst & Hoogendoorn (1988) who computed the flow in enclosures with  $A_H = 1$  and  $A_D = 1$  and 2 for Rayleigh numbers as large as  $10^{10}$ . However, we note that their finite difference meshes were very coarse ( $45 \times 45 \times 20$ ) and they made use of a  $k - \varepsilon$  turbulence model. As will be seen later, it appears highly unlikely that such models, which were developed for high Reynolds number flows, are applicable to low Reynolds number natural convection turbulence. Comparisons obtained from resolved ( $Ra \leq 10^6$ ) laminar results at the symmetry, plane ( $z = 0$ ) show that peak minimum and maximum velocities and heat transfer rates are within 10% of values obtained from two-dimensional calculations. Average values of the Nusselt number actually agree within 2%. In general the agreement is better at the larger values of Rayleigh numbers. Thus the major conclusion is that at high Rayleigh numbers, three-dimensional effects are insignificant in the bulk of the flow field. The exceptions are the regions near the end walls.

## 5. Cavities with $A_H \leq 1$

Cavities with small aspect ratios have not received as much attention as those with aspect ratios larger than unity. Furthermore, most analytical and experimental studies conducted to date on this problem have been in the Boussinesq or near-Boussinesq limit.

In exemplary studies, the problem was treated by Cormack *et al* (1974a, 1974b) and Imberger (1974) analytically, numerically, and experimentally, respectively.

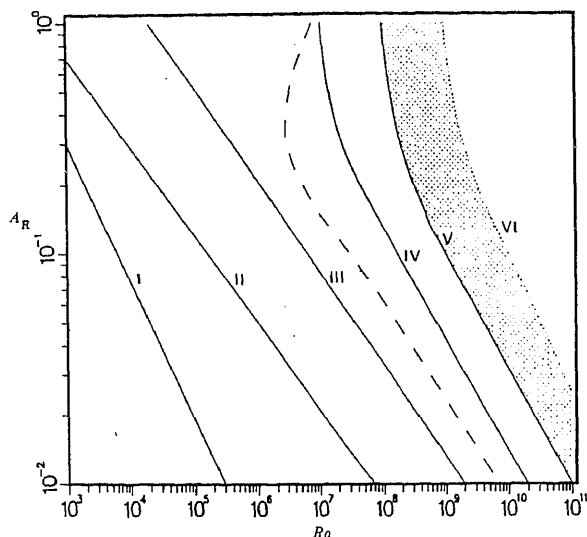


Cormack *et al* (1974a) presented an asymptotic solution to the laminar problem. This solution is valid in the aspect ratio limit of  $A_H \rightarrow 0$  for fixed, though arbitrary values of the Rayleigh and Prandtl numbers, and is assumed to consist of two distinct regions: a parallel flow in the central core (first obtained by Hart 1972) and a non-parallel flow confined within a distance of order  $H$  from the end walls. Cormack *et al* (1974b) developed an algorithm for the numerical solution of the flow in the cavity. They found their results to be in good agreement with the parallel flow solution obtained by Cormack *et al* (1974a) provided  $A_H \leq 0.1$  and  $Ra^2 A_H^3 \leq 10^5$ . In addition, the solutions show a parallel flow transition between the asymptotic limit of Cormack *et al* (1974a) and the boundary layer limit ( $A_H$  fixed and  $Ra \rightarrow \infty$ ) of Gill (1966). Imberger (1974) presented experimental results for water with  $A_H = 0.01$  and  $0.019$  in the range  $1.31 \times 10^6 \leq Ra \leq 1.11 \times 10^8$ . Most of the flow features indicated by the numerical work were qualitatively observed in the experimental work. When  $Ra^2 A_H^3$  becomes nearly  $10^{11}$ , experiments indicate that the mid-height of the cavity becomes an isotherm and there is a slow central circulation throughout the entire shallow cavity. We note that while  $A_H$  can be decreased by increasing the length of the cavity without varying the Rayleigh number, often to change the aspect ratio in an experiment the height is decreased thus forcing a large increase in temperature difference to keep the Rayleigh number fixed. Use of this experimental procedure to verify numerical Boussinesq results has limited validity since the numerical solutions do not account for property variations.

Bejan & Tien (1978) and Bejan (1980) extended the asymptotic analysis of Cormack *et al* (1974a) to include both  $Ra \rightarrow 0$ ,  $A_H$  finite, and  $A_H \rightarrow 0$ ,  $Ra$  finite. They also developed a Nusselt number correlation in the boundary layer regime and an empirical correlation including both limits. Bejan *et al* (1981) presented experimental results for water with  $A_H = 0.0625$  and  $2 \times 10^8 < Ra < 2 \times 10^9$ . They showed that, contrary to the assumption of Bejan & Tien (1978), for  $Ra^{1/4} A_H > 1$ , the core flow is non-parallel and is dominated by horizontal intrusions flowing along the two insulated horizontal walls and embracing a practically stagnant and thermally stratified fluid. In addition they observed weak counterflow. Finally, by statistically analysing previously published experimental work, Ostrach (1982), in a recent review paper, shows that the heat transfer varies with  $Ra^{0.3}$  for  $A_H < 0.1$  and  $Ra > 5 \times 10^5$ . He notes that the variation with  $Ra^{0.2}$  obtained by Bejan & Tien (1978) is incorrect since they assumed that the core flow was parallel in matching with the boundary layers in the end regions.

Paolucci & Chenoweth (1988) improved the understanding of the small aspect ratio problem. They emphasize non-Boussinesq effects arising from property variations due to large temperature differences. In addition, they extend the numerical solution in the Boussinesq regime to Rayleigh numbers larger than those previously available. Numerical solutions of the transient Navier-Stokes equations are used to generate laminar steady-state results. The use of the transient form of the equations was necessary to obtain oscillatory stability results. Their results are summarized in figure 3 where the various flow regions are classified according to different velocity field behaviours. This figure is the result of many calculations and is given in  $A_H - Ra$  space for  $Pr = 0.71$  and two values of  $\varepsilon$ ; the solid lines are for  $\varepsilon \rightarrow 0$  and the dashed line is for  $\varepsilon = 0.6$ . The numbered solid lines bound the different flow

To the left of I the core flow is parallel. Cormack *et al* (1974a), and the exact solution for this flow. It can be easily shown that in this region horizontal velocity is  $u_m = \pm Ra/72\sqrt{3}$  and its location is at  $y$  the region between I and II, parallel flow still exists in the core



**Figure 3.** Flow regions dependence on  $A_H$ ,  $Ra$  and  $\varepsilon$  for  $Pr = 0.71$ .

of the maximum velocity decreases below  $u_m$ , even though its location remains the same. Cormack *et al* (1974a) as well as Bejan & Tien (1978) give formulas for the variation of  $u_m$  which depend on aspect ratio and Rayleigh number. The flow remains unicellular between II and III, but the core flow becomes non-parallel and boundary layers exist next to the active walls. Thus all regions to the right of II are considered to be in the boundary layer regime. In the region between III and IV, secondary flows develop in the top-left and bottom-right (emerging) corner regions. These secondary flows separate and elongate as the Rayleigh number is increased and eventually grow to occupy most of the core region. The flow remains steady in the region between IV and V, but a weak tertiary flow develops in the centre of the core region. The rotation of this cell is the same as those of the primary and secondary flows. Finally to the right of V the flow becomes unsteady and eventually becomes turbulent to the right of VI. The unsteady flow is in the form of oscillations near the active walls. These oscillations, and resulting turbulence, are localized near the endwalls and emerging corners, and do not penetrate the central core region. We remark that throughout the entire boundary layer regime, the temperature in the central core region is approximately linearly stratified with horizontal isolines.

For  $\varepsilon = 0.6$  the flow is unicellular approximately up to line III. However, in contrast to the Boussinesq limit, no parallel flow was observed for this case which has the character of either that to the left or right of I. Due to property variations, the flow on the cold wall enters the boundary layer regime long before II, while that at the hot wall makes the transition significantly to the right of II. Similarly, secondary flows develop near the emerging corners at the cold end significantly before III and at the hot end much later than III. A most important result is that at the dashed line the cold wall boundary layer becomes unsteady. This same transition corresponds to line V in the Boussinesq limit. Note that there is very large reduction in the critical Rayleigh number with increasing  $\varepsilon$ . In contrast, the hot wall layer remains steady well past line V. Similarly, there is a large  $Ra$  range where the flow at the cold wall end becomes turbulent, but the hot wall region and the core flow remain laminar since the turbulence is localized and does not penetrate very far into the central core.

Contrary to the results of Shiralkar *et al* (1981), and Tichy & Gadgil (1982), Paolucci & Chenoweth (1988) find the presence of weak reverse flows in the core region for high  $Ra$  in the boundary layer regime, in agreement with the experimental results of Bejan *et al* (1981) and Al-Homoud & Bejan (1979). In addition, for still higher  $Ra$ , Paolucci & Chenoweth (1988) observe this flow bifurcating and leading to a tertiary flow consisting of a central cell having the same rotation as the primary and secondary cells. This last result has not been previously observed either experimentally or numerically.

In the non-Boussinesq regime, the velocity and temperature fields results of Paolucci & Chenoweth (1988) show significant dependence on  $\varepsilon$  especially near the side walls. As a result the well-known parallel flow solution, accurate in the core of the cavity for  $\varepsilon \ll 1$ , does not exist. For higher Rayleigh numbers, Paolucci & Chenoweth (1988) generalize the well-known analytical boundary layer solution of Gill (1966) to the case of arbitrary  $\varepsilon$ . Their solution and numerical results show that the cold-wall boundary layer is much thinner than the corresponding layer in the Boussinesq limit. As a result, the critical Rayleigh numbers for stationary and oscillatory instabilities are lowered with increasing temperature difference and are governed by the cold wall. In contrast to the high aspect ratio problem (see Chenoweth & Paolucci 1986), they find that the heat transfer and pressure also depend strongly on  $\varepsilon$ . Although the average Nusselt number is almost independent of  $\varepsilon$  in the boundary layer regime, this is not the case for lower Rayleigh numbers. Furthermore, the largest pressure change occurs in this same low Rayleigh number region. Both effects are physically related to the large departures from the parallel flow solution valid in the Boussinesq limit. These results show that there can be considerable risk if results obtained from the well-established Boussinesq limit are extrapolated for use where large temperature differences exist.

As in the case where  $A_H > 1$ , when the Prandtl number is of order unity or larger, the above results change little (see Quon 1972). However, when the Prandtl number is small, substantial changes appear. Hart (1983) and more recently Drummond & Korpela (1987) show that for  $Pr < 0.12$ , and aspect ratios less than the same value, a parallel flow core will exist up to  $Gr \approx 8000$  where secondary flow in the form of stationary transverse cells is seen to appear. These cells propagate out from the ends as an imperfect bifurcation. Spacing of the cells in the cavity has been shown to depend on  $Gr$ ,  $Pr$  and  $A_H$ . Cases were found for which new cells form and grow between existing cells as the space permitted. Other cases show cells merging and splitting. The effect of increasing  $Pr$  is to stabilize the flow so that for  $Pr > 0.12$  with  $A_H = 1/8$  and  $1/10$  no secondary motions were found.

## 6. Transition of boundary layer flow

The majority of the work dealing with differentially heated cavities has been concerned with steady-state laminar flow. Yet in many of the fields of application, the flow is unsteady and possibly turbulent. Since many variables of engineering interest depend strongly on the flow regime, it is essential to understand the different physical processes responsible for the conversion of an initially laminar flow to a turbulent one. With increasing Rayleigh number the problem becomes stiff owing to a decrease in boundary layer thickness. As a result, there has been very little numerical work performed in this area.

The study of stability of the natural convection flow in the differentially heated cavity presents a major difficulty: the non-existence of an exact analytical solution in the general case. Studies have proceeded by extrapolating results obtained in studying limiting cases where solutions can be obtained. These are cases that we have encountered earlier and are known as the conduction limit, the buoyancy layer, and the separate boundary layer regime. The latter case is characterized by the presence of a stable vertical temperature stratification. Such stability studies have been conducted by Gill & Davey (1969) for the buoyancy layer and Vest & Arpaci (1969) and Bergholz (1978) for separated boundary layers, on assuming two-dimensional perturbations. In these studies, the fluid is assumed to be linearly stratified. It was found that below a certain stratification level the instability was stationary, while above this critical stratification level it was oscillatory. Korpela (1974) studied the influence of the Prandtl number on the stability of the conduction regime. We note here that in order to relate these results to the closed cavity, the stratification level needs to be empirically related to  $A_H$  and  $Ra$ . Furthermore, the stratification is applied uniformly in the width of the cavity up to and including the active walls. Obviously this represents an approximation to the actual situation, and the extent of applicability to the differentially heated cavity can only be judged by comparison to full numerical studies of stability. In addition, Paolucci & Chenoweth (1989) show that their results applied to finite cavities (using a thermal stratification parameter near unity) yield a critical wavelength of the vertical wall boundary layer  $\lambda_c \sim A_H/Pr^n$ , where  $n \leq 1/2$ ; thus the applicability of their results to finite cavities is questionable for small Prandtl numbers since Paolucci & Chenoweth (1989) find  $\lambda_c \approx 0.3 A_H$  for  $Pr = 0.71$ , where  $n \approx 0$ . Lastly, these analyses cannot yield any possible instabilities due to the presence of the horizontal walls. Iyer (1973) showed that two-dimensional transverse waves are the most unstable in the buoyancy layer, thus indicating that a two-dimensional assumption is not unreasonable. However, it is not clear that this assumption will remain valid in the nonlinear regime. Patterson & Imberger (1980) were the first to propose a classification of types of transition regimes that one could encounter within the cavity. Depending on whether the Prandtl number is larger or smaller than unity, the authors discuss the time to establish steady solutions, when they exist, and the diverse states in which the flows evolve to the stationary solutions. They also give conditions for the existence of the separate boundary layer regime. Patterson & Imberger (1980) and Patterson (1984) gave a criterion for the presence of internal gravity waves that are observed in the core of the cavity when  $Pr \geq 1$  and  $A_H \leq 1$ . The presence of these gravity waves in the cavity is also discussed by Yewell *et al* (1982), Ivey (1984), and Thorpe (1968) who studied stationary gravity waves in fluids in the presence of continuous and discontinuous stratifications. Patterson & Imberger (1980) concluded that cavity-scale internal wave activity is due to a "pile up" of the horizontal intrusions at the far ends. Ivey (1984) performed experiments in a square cavity at Rayleigh numbers of the order of  $10^9$  using water as the working fluid. He emphasized the importance of the inertial effect of the flow and his results show that damped oscillations arise from internal hydraulic jumps caused by the turning of the vertical boundary layers. Since the source is localized, he further concludes that due to rapid attenuation, their presence could not be felt throughout the cavity as Patterson & Imberger (1980) suggest. The numerical results of Chenoweth & Paolucci (1986) seem to be in agreement with Ivey's (1984) conclusions; furthermore, they also suggest that for low aspect ratios the "hydraulic" jumps are responsible for the first transition to time-dependent flow.

Le Quéré & Alziary de Roquefort (1985) used a semi-implicit Chebyshev pseudo-spectral method to examine the oscillatory approach to steady-state of the average Nusselt number for  $A = 1$ ,  $Pr = 0.71$ , and  $Ra = 10^7$  and  $4 \times 10^7$ . For  $Ra = 4 \times 10^7$  they note the presence of detached regions near the departing corners at steady-state. These regions were also present at  $Ra = 10^7$ , but in this case they did not persist to steady-state. Subsequently, Le Quéré & Alziary de Roquefort (1986, pp. 1532–37) computed the first transition to periodic flow for  $A_H$  as low as 2 and concluded that in all cases the time-dependent periodic motion is a result of wall boundary layer instability. However, for  $A_H < 3$  they did note the presence of separated flow regions along the horizontal walls which remained when the flow was observed to be statistically stationary. Haldenwang (1986), also using a semi-implicit Chebyshev pseudo-spectral method, computed the solution for  $A_H = 1$ ,  $Pr = 0.71$ , and  $Ra = 10^6$ ,  $10^{6.5}$ ,  $10^7$ ,  $10^{7.5}$ ,  $10^8$ , and  $10^{8.5}$ . He concluded that: regions of reverse flow on the horizontal walls are present for  $Ra \geq 10^{7.5}$ ; the flow becomes oscillatory for  $Ra$  between  $10^8$ , and  $10^{8.5}$  where two fundamental frequencies were observed; and that these two frequencies, neither of which is in continuity with the one observed at smaller Rayleigh numbers, are first observed in the stable solution at  $Ra = 10^8$ .

Paolucci & Chenoweth (1989) address the oscillatory approach to steady-state and the transitions from steady-state to turbulence via two-dimensional direct numerical simulations. Their results clarify the basic mechanism of steady and unsteady oscillatory motion. Simulations are performed for a Boussinesq fluid with  $Pr = 0.71$ ,  $1/2 \leq A_H \leq 3$ , and a wide range of Rayleigh numbers. The resulting accuracy of their results is demonstrated by the excellent agreement with the results of Le Quéré & Alziary de Roquefort (1986), and Haldenwang (1986). Their restriction to two spatial dimensions precludes possibly important three-dimensional nonlinear effects due to vortex stretching. But even within their limitations, imposed by present day computers, the simpler model is of interest in providing insight into the physical mechanisms which drive the convective dynamics from laminar to turbulent flows. Furthermore, while it is an accepted fact that the laminar flow is inherently two-dimensional (see Eckert & Carlson 1961), some experiments indicate that even the resulting turbulent flow is dominated by two-dimensional structures (e.g. Elder 1965b). The goal behind their numerical experiments was to study the transitions to various time-dependent flows. With increasing Rayleigh number the onset of periodic flow is calculated for various aspect ratios. Power spectra of the temperature and velocity components are examined, and their dependence on the location probed is discussed. They are primarily concerned with instabilities that precede turbulence rather than strongly turbulent flows, although they also look at the oscillatory approach to steady-state for high Rayleigh numbers. The study of the loss of stability to time-dependent flow by direct numerical simulation allows them to obtain solutions for large supercritical values of the Rayleigh number. All their simulations satisfy a criterion for the presence of internal wave activity similar to that of Patterson & Imberger (1980), but for  $Pr \leq 1$  and arbitrary  $A_H$ .

As a result of the numerous computations for various values of  $A_H$  and  $Ra$ , Paolucci & Chenoweth (1989) obtain the stability map displayed in figure 4. This figure is a more detailed and accurate stability map of the lower right corner of figure 2. The solid and dashed lines in the figure represent the critical Rayleigh numbers  $Ra_i$  due to internal waves and  $Ra_w$  due to the wall boundary layers respectively. To the left of the curves, perturbations are damped, while to the right they are amplified leading to oscillatory flow. With increasing Rayleigh number, the steady convection flow

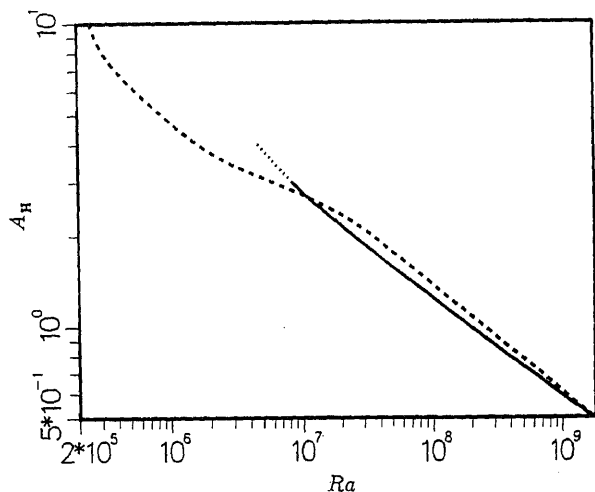


Figure 4. Critical Rayleigh numbers as functions of aspect ratio  $A_H$  for  $\varepsilon \rightarrow 0$  and  $Pr = 0.71$ : —,  $Ra_i$ ; ---,  $Ra_w$ .

becomes unstable with respect to time-dependent disturbances. In general, for  $A_H \leq 1/2$  and  $A_H \geq 3$  the first time-dependent instability is due to the boundary layers along the vertical walls. However, for  $1/2 < A_H < 3$  they first find periodic motion due to internal waves near the departing corners, then quasi-periodic motion (arising from wall boundary layers), aperiodic motion with complex regularity, and finally turbulent motion. They discuss at length the physical mechanisms responsible for the dynamic behaviours. Their results clearly show that for  $\varepsilon \rightarrow 0$ ,  $Pr = 0.71$  and  $1/2 < A_H < 3$  internal waves near the departing corners are indeed the source of low frequency oscillations as suggested by Ivey (1984), and not caused by a "pile up" of the horizontal intrusions at the far ends as argued by Patterson & Imberger (1980). While the high frequency oscillations are rapidly attenuated, the low frequency ones are not, so that their presence is felt throughout the cavity. The low frequency attenuation however is not in general as rapid as suggested by Ivey's (1984) experiment with water since their presence is felt throughout the cavity. In the range  $1/2 < A_H < 2$  the curves in figure 4 are described to a good approximation by

$$Ra_i = 1.93 \times 10^8 A_H^{-3.15}, \quad (28)$$

and

$$Ra_w = 2.70 \times 10^8 A_H^{-2.75}. \quad (29)$$

They also show that (28) and (29) are consistent with estimates obtained by the use of simple arguments and previous analyses.

In accord with Patterson & Imberger (1980) they show that within this region of parameter space the flow approaches steady-state conditions in an oscillatory fashion, although the source of the oscillations is different from that suggested by them. In agreement with Ivey's (1984) experimental results they find that the oscillatory behaviour is due to the inertia of the flow entering the interior of the cavity from the sidewall boundary layers (departing corners), which leads to a form of internal "hydraulic" jump when the Rayleigh number is sufficiently large. The onset and frequencies of the oscillatory instabilities are calculated and compared with available data. They present numerical experiments which exhibit repeated supercritical

branching leading to chaotic flow after a finite number of bifurcations. A sequence of instabilities quite similar to that described in their paper has been observed by Gollub & Benson (1980) in a laser-Doppler velocimetry study of Rayleigh-Bénard convection and also by Fenstermacher *et al* (1979) and Gorman & Swinney (1982) in the Couette-Taylor system. In particular, as seen in figure 5, with increasing Rayleigh number the time history and spectra show a periodic regime with a single fundamental frequency followed by a quasiperiodic regime with two fundamental frequencies, and then broadband noise components appear in the spectra; finally the amplitude of the sharp frequency components decrease. This particular sequence of bifurcations into chaos is further illustrated by the phase-space trajectories shown in figure 6. Thus, one principal finding of this work is that periodic flow is followed by only one additional distinct dynamical regime (quasi-periodic flow with two incommensurate frequencies) prior to the appearance of a chaotic regime. This result is consistent with the predictions of Newhouse *et al* (1978). In addition, even though they restricted the simulations to two spatial dimensions, their results compare favorably with the experimental results of Ivey (1984), and analytical/experimental studies of Thorpe (1968), and Keunecke (1970).

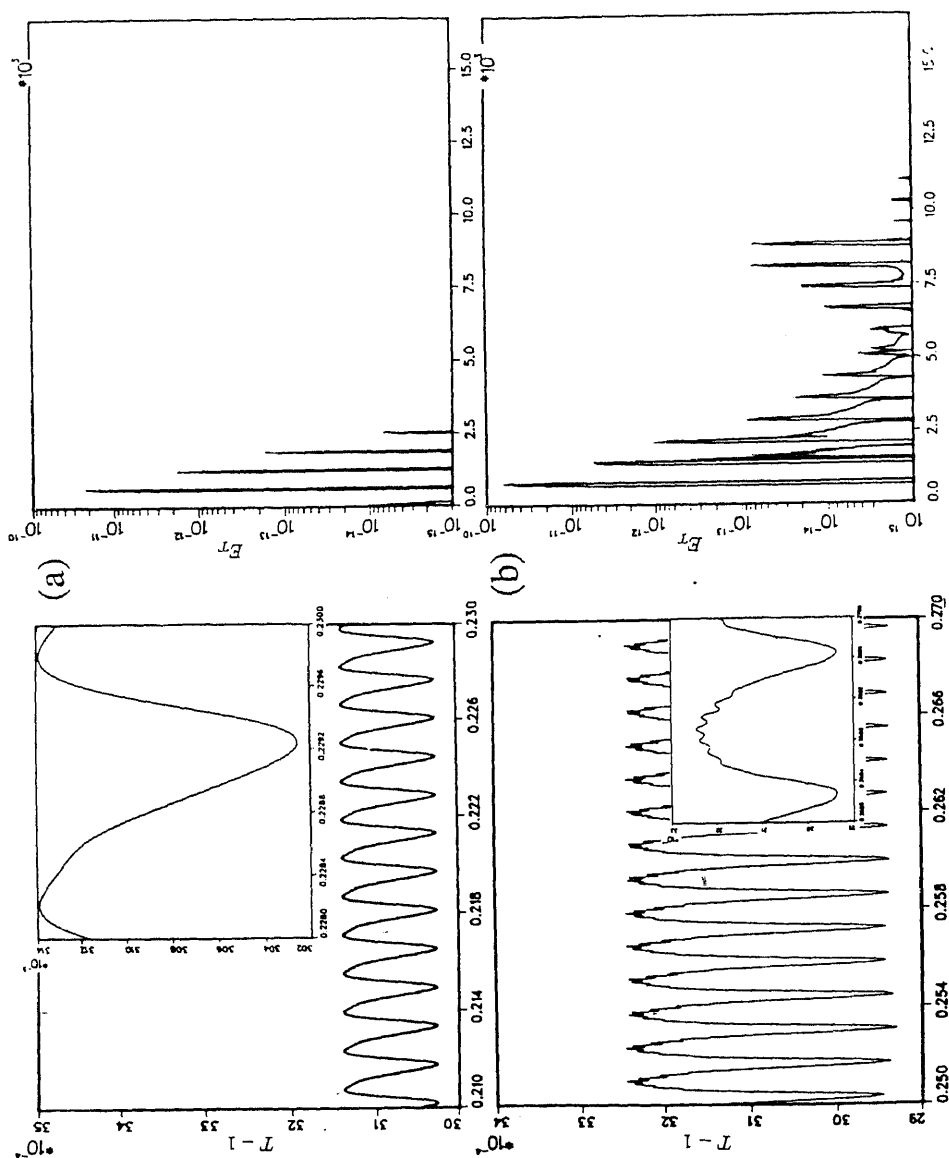
Briggs & Jones (1985) were the first to demonstrate experimentally different flows for the same values of parameters in a closed cavity having conducting horizontal walls. They demonstrated that in the Rayleigh number interval between  $6 \times 10^6$  and  $8 \times 10^6$  a hysteresis is present and a temporally oscillatory flow having one of three different frequencies can occur.

We remark that for  $A_H = O(1)$  and small Prandtl numbers, it becomes experimentally difficult to maintain adiabatic boundary conditions on the horizontal walls as  $Ra$  becomes large. Furthermore, in a physical experiment, often the Rayleigh number is increased by increasing the temperature difference across the cavity. It has been shown that both of these effects greatly modify the stability of the flow (Chenoweth & Paolucci 1986; Le Quéré & Alziary de Roquefort 1986).

## 7. Turbulence

While there has been substantial work devoted to the study of turbulent natural convection in the Rayleigh-Bénard problem where the gravitational vector is parallel to an imposed thermal gradient, relatively minor attention has been given to the case where gravity is orthogonal to the gradient. There are only a few publications on this subject and primarily in the Boussinesq regime. Furthermore, the published experimental data are insufficient to fully characterize the mechanisms responsible for momentum and heat transport in the thermal layers. In contrast, theoretical and experimental studies for the same problem but in the laminar regime are abundant. For example, see Elder (1965a), Cormack *et al* (1974a), De Vahl Davis & Jones (1983), Chenoweth & Paolucci (1986), and references therein.

In a vertical layer that is bounded by vertical isothermal surfaces having different temperatures and is thermally insulated at the ends, a circulatory flow is set up, ascending against the hot surface and descending at the cold surface. The flow in the cavity passes through several stages as the fluid flows along the active vertical walls. The flow near the entry corners of the boundary layers is initially laminar. It then passes through a transition region, and finally becomes turbulent. When statistical steady-state obtains, the space between the vertical boundary layers is filled by a





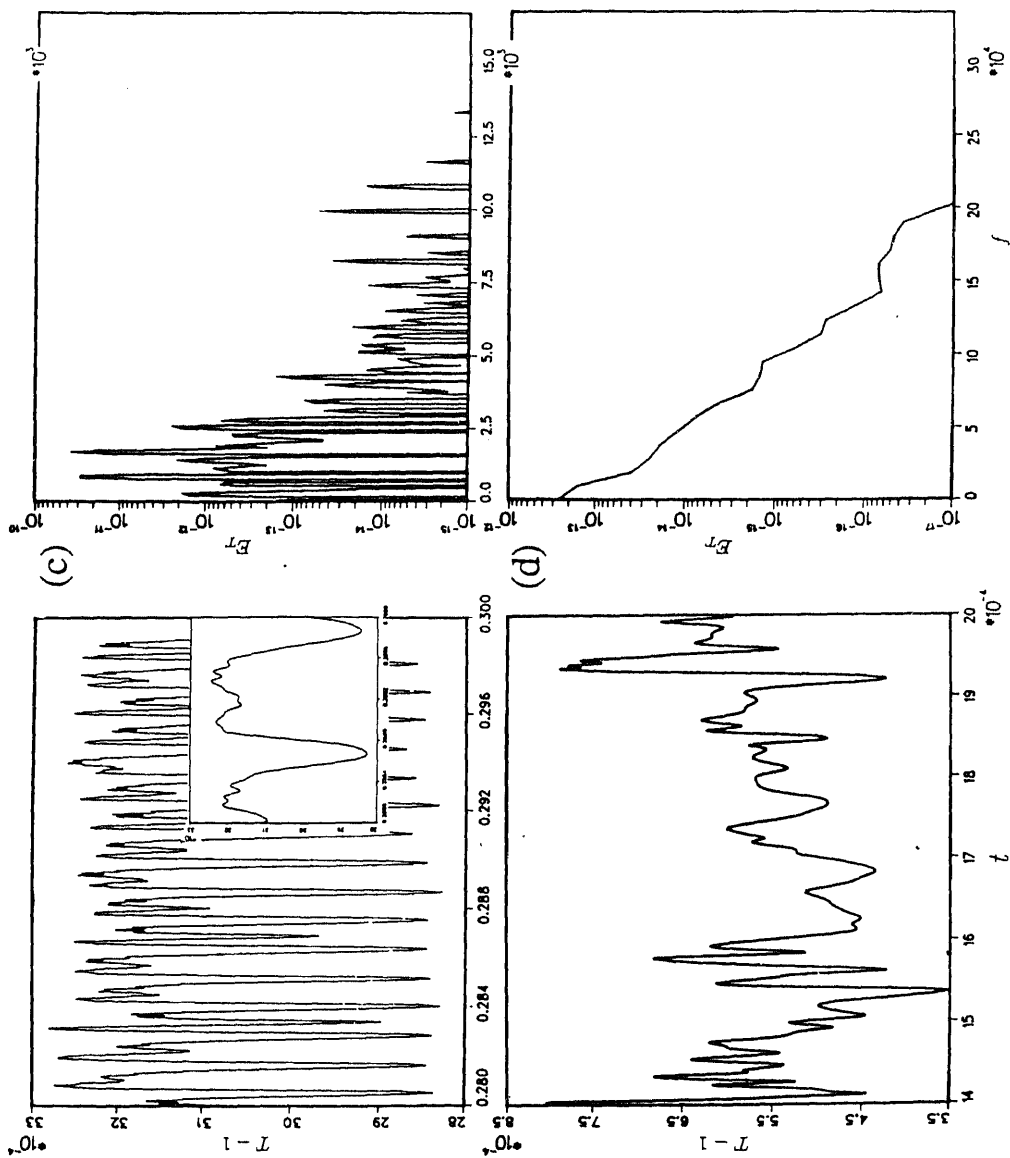


Figure 5. Time histories and power spectra of  $T-1$  for  $A_H = 1$ ,  $\varepsilon \rightarrow 0$  and  $Pr = 0.71$  at the point  $x_i = (1.032 \times 10^{-1}, 8.036 \times 10^{-1})$ ; (a)  $Ra = 2.0 \times 10^8$ , (b)  $Ra = 3.0 \times 10^8$ , (c)  $Ra = 4.0 \times 10^8$ , (d)  $Ra = 10^{10}$ .

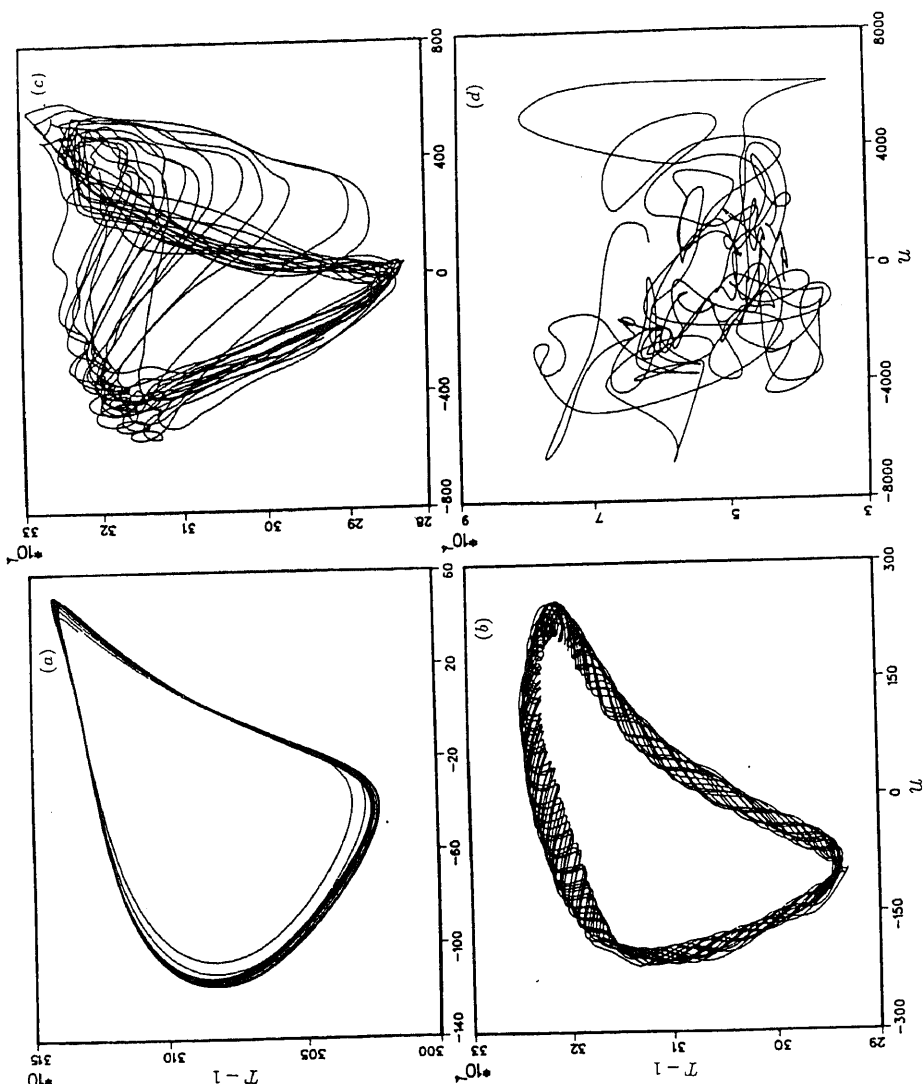


Figure 6. Phase-space trajectory of temperature versus velocity component  $u$  for  $A_H = 1$ ,  $\varepsilon \rightarrow 0$  and  $Pr = 0.71$  at the point  $x_i = (1.032 \times 10^{-1}, 8.036 \times 10^{-1})$ ; (a)  $Ra = 1.032 \times 10^{-1}$ , (b)  $Ra = 8.036 \times 10^{-1}$ , (c)  $Ra = 3.0 \times 10^8$ , (d)  $Ra = 4.0 \times 10^8$ .

virtually immobile stably stratified fluid executing low-frequency, low-velocity oscillations. The temperature away from the boundary layers increases linearly over a large part of the height of the layer.

One of the earliest experimental studies of this problem was performed by Mull and Reiher and has been discussed by Jakob (1949) and Batchelor (1954). Subsequently Mordchelles-Regnier & Kaplan (1963, pp. 94–111) have obtained some measurements and visualizations up to Rayleigh numbers of order  $10^{12}$ . These early works were followed by the classic experimental work of Elder (1965b), who measured primarily the temperature and velocity in the cavity using several fluids of high Prandtl number. He concluded that the occurrence of the turbulent wall layers in the cavity is independent of the aspect ratio and occurs in a similar manner to that on an isolated vertical plate. Subsequently, MacGregor & Emery (1969) experimentally obtained heat-transfer data for different Prandtl numbers and aspect ratios; Kutateladze *et al* (1972a, 1977, 1978) measured heat transfer, velocity and temperature means, their fluctuations, as well as the probability density distribution of temperature fluctuations; Cowan *et al* (1982, pp. 195–203) obtained overall and local heat-transfer data for different aspect ratios; Kirdyashkin *et al* (1983) and Kirdyashkin & Semenov (1984) measured the temperature means, fluctuations and their frequency spectra, the temperature kurtosis and skewness parameters, as well as other turbulence quantities; and Giel & Schmidt (1986, pp. 1459–64) obtained velocity and temperature means and fluctuations, as well as frequency spectra of temperature fluctuations.

Eckert & Carlson (1961), and Kutateladze *et al* (1972b, 1977) noted that the inner structure of the thermal boundary layer is locally similar to that of an isolated plate. More recently George & Capp (1979), using classical scaling arguments, presented a theory for turbulent natural-convection boundary layers next to heated vertical surfaces. They showed that the boundary layer must be treated in two parts: an outer region, in which the viscous and conduction effects are negligible, and an inner region, in which the mean convection terms are negligible. The inner layer, which is identified as a constant heat-flux layer in the sense that the total heat flux across the layer is independent of the distance from the wall, consists of two regions: a conductive and viscous sublayer, in which the temperature is linear and the velocity approximately linear next to the wall, and a buoyant sublayer outside of it, where the mean velocity and temperature profiles depend on the cube root and inverse cube root of distance from the wall, respectively.

Paolucci (1990) considers the nonlinear physics of turbulence numerically. More specifically, he performs a *direct* numerical simulation of the free convective flow in the cavity for  $A_H = 1$ ,  $\varepsilon \rightarrow 0$ ,  $Pr = 0.71$ , and  $Ra = 10^{10}$ . Although the initial conditions he used are non-random (quiescent and isothermal fluid), due to strong buoyancy the flow breaks up into an apparently *random* one. In general a flow which exhibits temporal as well as spatial randomness is identified as being turbulent.

The flow is again assumed to be two-dimensional. Although physical flows are three-dimensional, two-dimensional turbulence is often considered as a first approximation in many cases (e.g. turbulent flows submitted to a stable stratification). Furthermore, it appears from numerous experiments dealing with this problem (e.g., Giel & Schmidt 1986, pp. 1459–64), and the related problem of natural convection from a heated vertical plate (e.g., Lochet *et al* 1983), that three-dimensionality in this class of flows has a very small influence if the depth dimension is not too small. In this connection, it is noted that the inertial sub-range of two-dimensional turbulence is quite different from that in three-dimensions (Kraichnan 1967; Lilly 1969). In

particular, in inviscid isothermal flow the cascade of energy in two dimensions is mainly toward low wave number in the inertial range, and becomes weak or non-existent at higher wave numbers where only vorticity variance or enstrophy is cascaded in the usual sense in the inertial range.

Paolucci's (1990) results show that convection flow in a vertical layer for Rayleigh numbers in the turbulent regime passes through several stages. In the lower part of the cavity, next to the heated wall, the boundary layer flow has laminar character. This region is followed by transition and finally by a turbulent boundary layer. The space between the vertical boundary layers is filled by a virtually immobile stably stratified fluid executing low-frequency, low-velocity oscillations. This central core is continually stirred by random buoyant elements of finite energy which are discharged into it.

In accord with the picture of George & Capp (1979), the simulation shows that the thermal and momentum boundary layers can each be characterized by three regions. Directly against the wall there are the conductive and viscous sublayers, where the heat flux is constant but the shear stress is not. Defining the time averaged inner-region temperature and vertical velocity at the mid-height of the cavity ( $y = A_H/2$ ) as

$$\theta = (\bar{T} - \bar{T}_m)/(T_h - \bar{T}_m), \quad V_\eta = \eta \bar{v}, \quad (30)$$

where  $\eta = (\beta_r g \Delta T_m L^3 / \alpha_r^2)^{-1/3}$  is a dimensionless thermal inner scale,  $\bar{T}_m = \bar{T}(1/2, y)$ , and  $\Delta T_m = T_h - \bar{T}_m$ , these regions are then given by

$$\theta = 1 - 0.118(x/\eta) \quad \text{for} \quad 0 \leq x/\eta < 3.3, \quad (31)$$

and

$$V_\eta = \text{Pr}^{-1} [3.89(x/\eta) - \frac{1}{2}(x/\eta)^2 + (0.118/6)(x/\eta)^3], \quad \text{for} \quad 0 \leq x/\eta < 1.5. \quad (32)$$

In these regions a monotonic rise in the intensity of fluctuations is seen. These regions are followed by the thermal and momentum buoyant sublayers whose variations with distance from the wall are given by

$$\theta = 2.406(x/\eta)^{-1/3} - 0.943, \quad \text{for} \quad 3.3 < x/\eta < 13, \quad (33)$$

and

$$V_\eta = 13.63(x/\eta)^{1/3} - 8.77, \quad \text{for} \quad 1.5 < x/\eta < 3.3. \quad (34)$$

Thermal energy transferred by conduction accumulates in the thermal sublayer giving rise to fluctuations which constitute a considerable part of the heat transfer. Note that for  $\text{Pr} = 0.71$  the momentum buoyant sublayer is fully within the conductive sublayer. Exterior to the mean vertical velocity maximum the simulation results shows that the velocity distribution is self-similar and given by

$$\bar{v}/\bar{v}_m = \exp[-0.947[(x - x_m)/\delta_{\bar{v}}]^{1.432}], \quad (35)$$

where  $\bar{v}_m = 2.159 \times 10^4$  and  $x_m = 4.649 \times 10^{-3}$  are the maximum vertical velocity at the mid-height and the location away from the hot wall at which the maximum is found, and  $\delta_{\bar{v}} = \int_0^{1/2} (\bar{v}/\bar{v}_m) dx = 1.214 \times 10^{-2}$  is the momentum boundary layer thickness. In this strong mixing region, wave-like structures are superimposed on the mean motion. As elements of these structures accelerate out of the region, a local reduction of the thermal energy is observed. The elements move sufficiently rapidly

out of this highly intermittent region for molecular processes to be negligible. The external part of the thermal boundary layer is characterized by a small value of excess mean temperature.

It is seen that the stable stratification of the fluid outside the boundary layers significantly affects the nature of the flow, but has no effect on the heat transfer and momentum in the conductive and viscous sublayers. In the core of the cavity it is found that the temperature is approximately linearly stratified with a slope of

$$S = (A_H/2\varepsilon)(d\bar{T}/dy) = 0.38. \quad (36)$$

This result appears to be *independent* of  $A_H$ ,  $Pr$ , and most surprisingly  $Ra$  since it is in good agreement with the experimental results obtained over a wide range of Rayleigh numbers by Elder (1965) who obtained  $S = 0.3 - 0.4$ , for  $10 < A_H < 30$ , and  $Pr = 7$ , Kutateladze *et al* (1972a) who obtained  $S = 0.33$  for  $A_H = 22$  and  $Pr = 13.2$ , and Kutateladze *et al* (1977, 1978) who obtained  $S = 0.36 \pm 0.04$  for  $8 < A_H < 26$  and  $Pr \approx 16$ . In addition, using the statistically steady mean temperature gradient and averaging over the height of the cavity, he showed that the average Nusselt number can be written as

$$Nu = 0.046 Ra^{1/3}. \quad (37)$$

The agreement with the experimental values of Elder (1965b), MacGregor & Emery (1969), Kutateladze *et al* (1977), Cowan *et al* (1982), and Kirdyashkin *et al* (1983), is excellent, and, as with the value of  $S$  in (37), the constant 0.046 appears to be independent of the Prandtl number and aspect ratio, since the experimental results were obtained for  $1 \leq Pr \leq 10^3$  over the aspect ratio range  $1 < A_H < 61$ .

In comparing his predictions with available experimental evidence, many of Paolucci's (1990) results could be substantiated either directly or upon replotting data available in the literature. In particular, the heat-transfer law, the vertical stratification in the core, the viscous and conducting sublayers, the existence of momentum and thermal buoyant sublayers, and the self-similar distribution of the outer boundary layer region are in good quantitative agreement with available data. Many of the features observed in the flow such as the sinusoidal mode of instability, the internal jumps in the departing corners, and the "hook" like structures have been observed experimentally. However, the validity of many other results can only be verified when more experimental data become available.

Two important points can be made from the direct simulation results. First, there is now strong evidence that the prevailing theories of heat and mass transfer which assume relationships for friction and heat transfer, similar to those in forced boundary layer flow, are not applicable to natural-convection turbulence. Second, the results should be helpful in constructing simplified turbulence models for natural-convection flows in the future.

## 8. Conclusions

We have presented a brief review of recent work on the natural convection flow in enclosures. The prototypical configuration used in studying these flows is that of a vertical rectangular enclosure with lateral heating. The buoyancy forces in this review are considered to result exclusively from temperature differences applied across two

vertical walls. The natural convection flows discussed have been primarily in enclosures in which  $A_D \rightarrow \infty$ , and for both large and small vertical aspect ratios  $A_H$ .

It is apparent from our discussion that the subject of natural convection in two-dimensional large and small vertical aspect ratios and in the Boussinesq limit has a rather long history. Most of the elementary aspects of these flows in the laminar regime are now fairly well understood. More research is needed in the following areas.

- (1) Laminar and transitional flows in cavities where  $A_H = O(1)$ , and  $A_D = O(1)$  in both the Boussinesq and non-Boussinesq regimes. Very limited work has been done in this area computationally by Mallinson & De Vahl Davis (1977) and Fusegi *et al* (1991), and experimentally by N'Dame (1992) who recently gave evidence of a possible low-frequency three-dimensional instability for  $A_D \geq 1.37$ .
- (2) More complete study of the multiple solution regimes first found by Briggs & Jones (1985). Some of this work has been started by Briggs & Jones (1989), Le Quéré & Alziary de Roquefort (1988), Le Quéré (1990), and Penot *et al* (1990, pp. 417–22).
- (3) A more accurate and complete study of natural convection turbulence. This should be done using a more accurate method than that used by Paolucci (1990) to verify many of his results, should be extended to three-dimensions to eliminate many of the questions that have been raised about that work, and should be further extended to the full non-Boussinesq regime. In addition, and most importantly, these direct simulation results should be used to generate simple models to replace current models that are fundamentally based on high Reynolds number forced convection theories that assume classical relationships between friction and heat transfer.

## References

- Al-Homoud A A, Bejan A 1979 Experimental study of high Rayleigh number convection in a horizontal cavity with different end temperatures. Tech. Rep. CUMER-79-1, University of Colorado, Boulder, Dept. of Mech. Eng.
- Arakawa A 1966 Computational designs for long term numerical integration of the equations of fluid motion – Part 1. Two dimensional incompressible flow. *J. Comput. Phys.* 1: 119–143
- Batchelor G K 1954 Heat transfer by free convection across a closed cavity between vertical boundaries at different temperatures. *Q. Appl. Math.* 12: 209–233
- Bejan A 1979 Note on Gill's solution for free convection in vertical enclosure. *J. Fluid Mech.* 90: 561–568
- Bejan A 1980 A synthesis of analytical results for natural convection heat transfer across rectangular enclosures. *Int. J. Heat Mass Transfer* 23: 723–726
- Bejan A, Al-Homoud A A, Imberger J 1981 Experimental study of high Rayleigh number convection in a horizontal cavity with different end temperatures. *J. Fluid Mech.* 109: 283–299
- Bejan A, Tien C L 1978 Laminar natural convection heat transfer in a horizontal cavity with different end temperatures. *J. Heat Transfer* 100: 641–647
- Bergholz R F 1978 Instability of steady natural convection in a vertical fluid layer. *J. Fluid Mech.* 84: 743–768
- Briggs D G, Jones D N 1985 Two-dimensional periodic natural convection in a rectangular enclosure of aspect ratio one. *J. Heat Transfer* 107: 850–854
- Briggs D G, Jones D N 1989 Periodic two-dimensional cavity flow: effect of linear horizontal thermal boundary conditions. *J. Heat Transfer* 111: 86–91
- Catton I 1978 Natural convection in enclosures. In *Heat Transfer 1978* (Washington, DC: Hemisphere) vol. 6

- Chenoweth D R, Paolucci S 1985 Gas flow in vertical slots with large horizontal temperature differences. *Phys. Fluids* 28: 2365–2374
- Chenoweth D R, Paolucci S 1986 Natural convection in an enclosed vertical air layer with large horizontal temperature differences. *J. Fluid Mech.* 169: 173–210
- Cormack D E, Leal L G, Imberger J 1974a Natural convection in a shallow cavity with differentially heated end walls. Pt. 1, Asymptotic theory. *J. Fluid Mech.* 65: 209–229
- Cormack D E, Leal L G, Seinfeld J H 1974b Natural convection in a shallow cavity with differentially heated end walls. Pt. 2, Numerical solutions. *J. Fluid Mech.* 65: 231–246
- Cowan G H, Lovegrove P C, Quarini G L 1982 Turbulent natural convection heat transfer in vertical single water-filled cavities. In *Heat Transfer 1982* (eds) U Grigull, E Hahne, K Stephan, J Straub (Washington DC: Hemisphere) vol. 2
- De Vahl Davis G, Jones I P 1983 Natural convection in a square cavity: a comparison exercise. *Int. J. Numer. Meth. Fluids* 3: 227–248
- Deardorff J W 1973 Three-dimensional numerical modelling of the planetary boundary layers. In *Workshop on micrometeorology* (Washington, DC: Am. Math. Soc.)
- Drummond J E, Korpela S A 1987 Natural convection in a shallow cavity. *J. Fluid Mech.* 182: 543–564
- Duxbury D 1979 *An interferometric study of natural convection in enclosed plane air layers with complete and partial central vertical divisions*. Ph D thesis, University of Salford, UK
- Eckert E R G, Carlson W O 1961 Natural convection in an air layer enclosed between two vertical plates with different temperatures. *Int. J. Heat Mass Transfer* 2: 106–120
- Elder J W 1965a Laminar free convection in a vertical slot. *J. Fluid Mech.* 23: 77–98
- Elder J W 1965b Turbulent free convection in a vertical slot. *J. Fluid Mech.* 23: 99–111
- Fenstermacher P R, Swinney H L, Gollub J P 1979 Dynamical instabilities and the transition to chaotic Taylor vortex flow. *J. Fluid Mech.* 94: 103–128
- Fusegi T, Hyun J M, Kuwahara K, Farouk B 1991 A numerical study of three-dimensional natural convection in a differentially heated cubical enclosure. *Int. J. Heat Mass Transfer* 34: 1543–1557
- George W K, Capp S P 1979 A theory for natural convection turbulent boundary layers next to heated vertical surfaces. *Int. J. Heat Mass Transfer* 22: 813–826
- Giel P W, Schmidt F W 1986 An experimental study of high Rayleigh number natural convection in an enclosure. In *Heat Transfer 1986* (eds) C L Tien, V P Carey, J K Ferrell (Washington: DC: Hemisphere) vol. 4
- Gill A E 1966 The boundary-layer regime for convection in a rectangular cavity. *J. Fluid Mech.* 26: 515–536
- Gill A E, Davey A 1969 Instabilities of a buoyancy-driven system. *J. Fluid Mech.* 35: 775–798
- Gollub J P, Benson S V 1980 Many routes to turbulent convection. *J. Fluid Mech.* 100: 449–470
- Gorman M, Swinney H L 1982 Spatial and temporal characteristics of modulated waves in the circular Couette system. *J. Fluid Mech.* 117: 123–142
- Graebel W P 1981 The influence of Prandtl number on free convection in a rectangular cavity. *Int. J. Heat Mass Transfer* 23: 125–131
- Gray D D, Giorgini A 1976 The validity of the Boussinesq approximation for liquids and gases. *Int. J. Heat Mass Transfer* 19: 545–551
- Haldenwang P 1986 Unsteady numerical simulation by Chebyshev spectral methods of natural convection at high Rayleigh number. *J. Fluid Mech.* 144: 389–401
- Hara T 1958 Heat transfer by laminar free convection about a vertical flat plate with large temperature difference. *Bull. JSME* 1: 251–254
- Hart J E 1972 Stability of thin non-rotating Hadley circulations. *J. Atmos. Sci.* 29: 687–697
- Hart J E 1983 Low Prandtl number convection between differentially heated end walls. *Int. J. Heat Mass Transfer* 26: 1069–1074
- Imberger J 1974 Natural convection in a shallow cavity with differentially heated end walls. Pt. 3, Experimental results. *J. Fluid Mech.* 65: 247–260
- Ivey G N 1984 Experiments on transient natural convection in a cavity. *J. Fluid Mech.* 144: 389–401
- Iyer P A 1973 Instabilities in buoyancy-driven boundary-layer flows in a stably stratified medium. *Boundary-Layer Meteorology* 5: 53–66
- Jakob M 1949 *Heat transfer* (New York: Wiley)
- Keunecke V K -H 1970 Stehende interne Wellen in rechteckigen Becken (Standing internal waves in rectangular tanks). *Deutsche Hydrographische Zeitschrift* 23: 61–79

- Kirdyashkin A G, Semenov V I 1984 Spectra of temperature fluctuations in a vertical layer with thermogravitational convection. *High Temperature* 21: 558–565
- Kirdyashkin A G, Semenov V I, Berdnikov V S, Gaponov V A 1983 Structure of the temperature field in a vertical layer with thermal gravitational convection. *High Temperature* 20: 750–757
- Korpela S A 1974 A study of the effect of Prandtl number on the stability of the conduction regime of natural convection in an inclined slot. *Int. J. Heat Mass Transfer* 17: 215–222
- Korpela S A, Gözüüm D, Baxi C B 1973 On the stability of the conduction regime of natural convection in a vertical slot. *Int. J. Heat Mass Transfer* 16: 1683–1690
- Kraichnan R H 1967 Inertial ranges in two-dimensional turbulence. *J. Phys. Fluids* 10: 1417–1423
- Kutateladze S S, Ivakin V P, Kirdyashkin A G, Kekalov A N 1977 Turbulent natural convection in a vertical layer. *High Temperature* 15: 458–464
- Kutateladze S S, Ivakin V P, Kirdyashkin A G, Kekalov A N 1978 Thermal free convection in a liquid in a vertical slot under turbulent flow conditions. *Heat Transfer – Soviet Res.* 10: 118–125
- Kutateladze S S, Kirdyashkin A G, Ivakin V P 1972a Turbulent natural convection on a vertical plate and in a vertical layer. *Int. J. Heat Mass Transfer* 15: 193–202
- Kutateladze S S, Kirdyashkin A G, Ivakin V P 1972b Turbulent natural convection at an isothermal vertical plate. *High Temperature* 10: 76–79
- Lankhorst A M, Hoogendoorn C J 1988 Three-dimensional numerical calculations of high Rayleigh number natural convective flows in enclosed cavities. In *Proc. 1988 National Heat Transfer Conf., ASME HTD-96* (New York: ASME) 3: 463–470
- Lauriat G 1980 Numerical study of natural convection in a narrow vertical cavity: an examination of high-order accuracy schemes. ASME Paper 80-HT-90
- Le Quéré P 1990 A note on multiple and unsteady solutions in two-dimensional convection in a tall cavity. *J. Heat Transfer* 112: 965–974
- Le Quéré P, Alziary de Roquefort T 1985 Computation of natural convection in two-dimensional cavities with Chebyshev polynomials. *J. Comput. Phys.* 57: 210–228
- Le Quéré P, Alziary de Roquefort T 1986 Transition to unsteady natural convection in a differentially heated cavity. In *Heat Transfer 1986* (eds) C L Tien, V P Carey, J K Ferrell (Washington, DC: Hemisphere) vol. 4
- Le Quéré P, Alziary de Roquefort T 1988 Sur l'existence de solutions périodiques multiples aux équations de Boussinesq. *C. R. Acad. Sci. Paris, II* 306: 681–687
- Le Quéré P, Masson R, Perrot P 1992 A Chebyshev collocation algorithm for 2D non-Boussinesq convection. *J. Comput. Phys.* 103: 320–335
- Lee T S, Son G H, Lee J S 1988 Numerical predictions of three-dimensional natural convection in a box. In *Proc. 1st KSME-JSME Thermal and Fluids Eng. Conf.* 2: 278–283
- Lee Y, Korpela S A, 1983 Multicellular natural convection in a vertical slot. *J. Fluid Mech.* 126: 91–121
- Leonardi E, Reizes J A 1979 Natural convection in compressible fluids with variable properties. In *Numerical methods in thermal problems* (eds) R W Lewis, K Morgan (Swansea, UK: Pineridge)
- Leonardi E, Reizes J A 1981 Convective flows in closed cavities with variable fluid properties. In *Numerical methods in heat transfer* (eds) R W Lewis, K Morgan, O C Zienkiewicz (New York: John Wiley)
- Lilly D K 1969 Numerical simulation of two-dimensional turbulence. *Phys. Fluids Supp.* II 12: 240–249
- Lochet R, Lemonnier D, Doan-Kim-Son 1983 Correlations en convection naturelle turbulente influence de la pression et de la nature du gaz. *Int. J. Heat Mass Transfer* 26: 1221–1227
- MacGregor R K, Emery A F 1969 Free convection through vertical layers—Moderate and high Prandtl number fluids. *J. Heat Transfer* 91: 391–401
- Mallinson G D, De Vahl Davis G 1977 Three-dimensional natural convection in a box: a numerical study. *J. Fluid Mech.* 83: 1–31
- Mordchelles-Regnier G, Kaplan C 1963 Visualization of natural convection on a plane wall and in a vertical gap by differential interferometry: transitional and turbulent regimes. In *Heat Transfer and Fluid Mechanics Institute* (Stanford: University Press)
- N'Dame A 1992 *Etude expérimentale de la convection naturelle en cavité: de l'état stationnaire au chaos*. Ph D thesis, Université de Poitiers, France



- Newhouse S, Ruelle D, Takens P 1978 Occurrence of strange axiom A attractors near quasi periodic flows on  $T^m$ ,  $m \geq 3$ . *Commun. Math. Phys.* 64: 35–40
- Ostrach S 1972 Natural convection in enclosures. In *Advances in heat transfer* (New York: Academic Press) vol. 8
- Ostrach S 1982 Natural convection heat transfer in cavities and cells. In *Heat Transfer 1982* (ed.) U Grigull, E Hahne, K Stephan, J Straub (Washington, DC: Hemisphere) vol. 1
- Paolucci S 1982 On the filtering of sound from the Navier–Stokes equations. Tech. Rep. SAND82–8257, Sandia National Laboratories Report
- Paolucci S 1990 Direct numerical simulation of two-dimensional turbulent natural convection in an enclosed cavity. *J. Fluid Mech.* 215: 229–262
- Paolucci S, Chenoweth D R 1988 Natural convection in shallow enclosures with differentially heated endwalls. *J. Heat Transfer* 110: 625–634
- Paolucci S, Chenoweth D R 1989 Transition to chaos in a differentially heated vertical cavity. *J. Fluid Mech.* 201: 379–410
- Patterson J C 1984 On the existence of an oscillatory approach to steady natural convection in cavities. *J. Heat Transfer* 106: 104–108
- Patterson J C, Imberger J 1980 Unsteady natural convection in a rectangular cavity. *J. Fluid Mech.* 100: 65–86
- Penot F, N'Dame A, Le Quéré P 1990 Investigation of the route to turbulence in a differentially heated cavity. In *Heat Transfer 1990* (ed.) G Hetsroni (Washington, DC: Hemisphere) vol. 2
- Peyret R, Taylor T D 1983 *Computational methods for fluid flow* (New York: Springer)
- Piacsek S A, Williams G P 1970 Conservation properties of convection difference schemes. *J. Comput. Phys.* 6: 362–392
- Polezhaev V I 1967 Numerical solution of a system of two-dimensional unsteady Navier–Stokes equations for a compressible gas in a closed region. *Fluid Dynamics* 2: 70–74
- Polezhaev V I 1968 Flow and heat transfer in laminar natural convection in a vertical layer. *Teplo i Massoobmen* 1: 631–640
- Quon C 1972 High Rayleigh number convection in an enclosure – A numerical study. *Phys. Fluids* 15: 12–19
- Roux B, Grondin J C, Bontoux P, De Vahl Davis G 1980 Reverse transition from multicellular to monocellular motion in vertical fluid layer. *Phys. Chem. Hydrodyn.* 3: 292–297
- Rubel A, Landis F 1970 Laminar natural convection in a rectangular enclosure with moderately large temperature differences. In *Heat Transfer 1970* (eds) U Grigull, E Hahne (Washington, DC: Hemisphere) vol. 4
- Shiralkar G S, Gadgil A, Tien C L 1981 High Rayleigh number convection in shallow enclosures with different end temperatures. *Int. J. Heat Mass Transfer* 24: 1621–1629
- Sparrow E M, Gregg J L 1958 The variable fluid-property problem in free convection. *Trans. ASME* 80: 879–886
- Thorpe S A 1968 On standing internal gravity waves of finite amplitude. *J. Fluid Mech.* 32: 489–528
- Tichy J, Gadgil A 1982 High Rayleigh number laminar convection in low aspect ratio enclosures with adiabatic horizontal walls and differentially heated vertical walls. *J. Heat Transfer* 104: 103–110
- Vest C M, Arpaci V S 1969 Stability of natural convection in a vertical slot. *J. Fluid Mech.* 36: 1–15
- White F M 1974 *Viscous fluid flow* (New York: McGraw-Hill)
- Yang K T 1987 Natural convection in enclosures. In *Handbook of single-phase convective heat transfer* (eds) S Kakac, R K Shah, W Aung (New York: Wiley-Interscience)
- Yang K T 1988 Transitions and bifurcations in laminar buoyant flows in confined enclosures. *J. Heat Transfer* 110: 1191–1204
- Yewell R, Poulikakos D, Bejan A 1982 Transient natural convection – Experiments in shallow enclosures. *J. Heat Transfer* 104: 533–538
- Yin S H, Wung T Y, Chen K 1978 Natural convection in an air layer enclosed within rectangular cavities. *Int. J. Heat Mass Transfer* 21: 307–315



## Flow transitions and pattern selection of the Rayleigh–Bénard problem in rectangular enclosures

D MUKUTMONI<sup>1</sup> and K T YANG<sup>2</sup>

<sup>1</sup>Adapco Ltd, Melville, NY 11747, USA

<sup>2</sup>Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

**Abstract.** In this paper, research efforts in the broad area of flow transitions of Rayleigh–Bénard convection in rectangular enclosures with sidewalls are reviewed. Numerical studies are given primary emphasis. However, experimental works that are relevant are described. Our current physical understanding of the transition phenomena as occurring in the Rayleigh–Bénard problem is critically reviewed. Two broad categories of transition are discussed. In the former, the transitions are temporal in nature, and mostly confined to small enclosures. In the latter, transitions are a result of change in spatial patterns. This phenomenon, known as pattern selection, is looked into for both small and intermediate enclosures.

**Keywords.** Rayleigh–Bénard problem; rectangular enclosure; flow transition; pattern selection.

### 1. Introduction

Rayleigh–Bénard convection is among the most heavily investigated physical problems for nearly a century. The enduring popularity of the problem with respect to a wide cross section of scientists and engineers is not surprising. The Rayleigh–Bénard problem is relevant to applications ranging from astrophysics, geophysics, atmospheric sciences and various disciplines of engineering. Many familiar and esoteric physical phenomena such as the imperceptible movement of the continental plates, the violent magnetic storms in the solar atmosphere, the destructive forces of a tropical cyclone are in essence a manifestation of the Rayleigh–Bénard convection for different geometries and parametric ranges. In an industrial context, Rayleigh–Bénard (henceforth abbreviated to RB at times) convection has applications in disciplines such as solar energy systems, energy storage, material processing and nuclear reactor systems.

RB convection forms a subclass of fluid flow problems that are known as buoyancy-driven flows or thermal convection. In these, fluid flow is induced by density differences that arise as a result of temperature differences. Thermal convection occurring in enclosures or cavities caused by a temperature gradient in the direction of the gravity vector is known as Rayleigh–Bénard convection. Other than its relevance to the

various disciplines of the physical sciences, the RB problem has been investigated for theoretical and fundamental reasons as well. The problem is governed by nonlinear coupled partial differential equations. It therefore serves as a paradigm of a nonlinear system. The model problem has been used to investigate the transition from laminar to turbulent flow. In another related issue, the physical model has been used to study the nonlinear selection and evolution of patterns. These two aspects of RB convection will be the subject of this paper.

More specifically, we review and highlight some of the computational work that has been done to study the phenomena of flow transition and pattern formation with respect to Rayleigh–Bénard convection. In the process, we will also include some of the relevant experimental work. This is important, since the only true test of the correctness of the computational model is reasonable agreement with experiments. The twin areas offer an exciting area of research with far-reaching implications.

## **2. Brief background of the Rayleigh–Bénard problem**

The Rayleigh–Bénard problem in its simplest form and one that was the earliest to be investigated is the so-called infinite layer case. In such a case, a layer of fluid is constrained between two infinite horizontal surfaces. The system is heated from below, i.e., the lower surface is at a higher temperature than the upper surface. The heated-from-below case is said to have an adverse temperature gradient because the fluid at the bottom will be lighter than the fluid at the top and this top-heavy arrangement is potentially unstable. When the temperature gradient is below a certain value, the natural tendency of the fluid to move, because of buoyancy, will be inhibited by its own viscosity and thermal diffusivity. Thus, the thermal instability will manifest itself only when the adverse temperature gradient exceeds a certain critical value.

The earliest experiments to demonstrate in a definitive manner the onset of thermal instability in fluids are due to Bénard (1900). The theoretical foundations for a correct interpretation of the phenomena are due to Rayleigh (1916). The phenomenon of thermal convection under an adverse temperature gradient is therefore known as the Rayleigh–Bénard convection in their honour. The non-dimensional adverse temperature gradient is known as the Rayleigh number. An introductory exposition and review of the infinite layer RB convection can be found in Chandrasekhar (1961). A more advanced and complete review of the infinite layer RB convection instability at higher Rayleigh numbers is given in Busse (1978). For a general discussion of transitions and bifurcations for buoyancy-driven enclosure flows, the reader is referred to Yang (1988).

Although it will be of immense benefit to the research community to review all current work done with regard to flow transitions in natural convection, it is felt that the scope is too vast to be included in just a single article. We therefore chose a more specialized topic and restricted ourselves to the numerical study of flow transitions, instabilities and bifurcations for the Rayleigh–Bénard problem in rectangular enclosures with sidewalls only. In the following sections, the basic RB problem is formulated and the various physical parameters discussed. This is followed by a general discussion of RB convection with respect to flow transitions primarily restricted to small enclosures and temporal transitions. Following that, we discuss RB convection restricted to spatial transitions in small and intermediate boxes. All along we emphasize the

numerical methods used by the researchers in their study and relevant experimental works. We conclude with a recommendation on future study.

### 3. Problem statement

The problem formulation described in this section was for most part used for most of the numerical work reviewed and in particular to all our recent investigations. RB convection is studied in a three-dimensional rectangular geometry. The geometry of the enclosure is shown in figure 1. The vertical walls are adiabatic. The bottom wall is heated and the top wall is cooled, both isothermally. The Boussinesq approximation is invoked. Consequently, all transport properties are assumed constant with the exception of the buoyancy term in the momentum equations, which is linearized. The governing equations are non-dimensionalized by suitable scales of the dependent and independent variables. The  $x$ ,  $y$  and  $z$  coordinates are non-dimensionalized by  $L$ , the enclosure height. The velocities, time and pressure were scaled by  $\alpha/L$ ,  $L^2/\alpha$  and  $\rho L^2/\alpha^2$  respectively. In this case,  $\alpha$  is the thermal diffusivity of the fluid. The temperature was normalized with respect to the top and bottom wall temperatures. The non-dimensionalized governing equations are the continuity, the Navier-Stokes and the energy equations that are listed below:

$$\nabla \bullet \mathbf{U} = 0, \quad (1)$$

$$(\partial u / \partial t) + \nabla \bullet (\mathbf{u} \mathbf{U}) = -(\partial p / \partial x) + \text{Pr} \nabla^2 u, \quad (2)$$

$$(\partial v / \partial t) + \nabla \bullet (\mathbf{v} \mathbf{U}) = -(\partial p / \partial y) + \text{Pr} \nabla^2 v + \text{RaPr} T, \quad (3)$$

$$(\partial w / \partial t) + \nabla \bullet (\mathbf{w} \mathbf{U}) = -(\partial p / \partial z) + \text{Pr} \nabla^2 w, \quad (4)$$

$$(\partial T / \partial t) + \nabla \bullet (\mathbf{T} \mathbf{U}) = \nabla^2 T. \quad (5)$$

The boundary conditions that are consistent with the adiabatic and isothermal walls

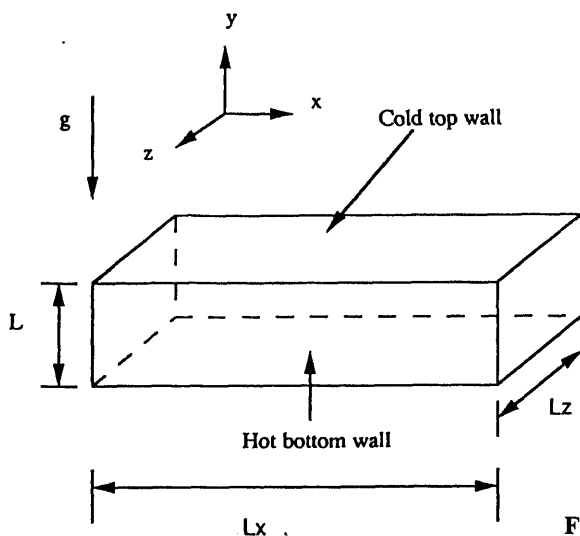


Figure 1. Geometry of the enclosure.

in a non-dimensional form are the following:

$$x = 0, A_x; 0 \leq z \leq A_z; 0 \leq y \leq 1; u = v = w = 0 (\partial T / \partial x) = 0, \quad (6)$$

$$z = 0, A_z; 0 \leq x \leq A_x; 0 \leq y \leq 1; u = v = w = 0 (\partial T / \partial z) = 0, \quad (7)$$

$$y = 0, 1; 0 \leq x \leq A_x; 0 \leq z \leq A_z; u = v = w = 0 \quad T = 0.5 - y. \quad (8)$$

The significant parameters of the formulation are the Rayleigh number  $Ra$  that represents the strength of the driving forces and the Prandtl number  $Pr$ . They are defined as follows;

$$Ra = g\beta\Delta TL^3/(v\alpha); \quad Pr = v/\alpha. \quad (9)$$

The two other parameters are the normalized dimensions of the rectangular enclosure in the horizontal directions or the geometrical aspect ratios  $A_x$  and  $A_z$ . For problems in which some of the assumptions are relaxed, there will be additional parameters. For instance if the Boussinesq approximation is not used, a parameter to quantify the departure from the Boussinesq approximation will arise. These additional parameters will be discussed on a case-by-case basis wherever applicable.

#### 4. Bifurcation to convective flow from initial conduction state

It has generally been known that for the given problem, as the Rayleigh number is increased, the system undergoes a series of bifurcations. By a bifurcation we mean that the flow undergoes a *qualitative* change in the flow and temperature field. For instance, an increase in the Rayleigh number and a corresponding decrease in the number of rolls would exemplify a flow bifurcation. On the other hand, if an increase in the Rayleigh number merely increases the velocities and heat transport without changing the flow pattern, no bifurcation would have occurred.

Below a certain critical Rayleigh number, there is no motion since the buoyant forces cannot overcome the viscous forces. Davis (1967) and Catton (1970) calculated the critical Rayleigh number for the onset of convection by a linear stability analysis of the motionless conduction state. The eigenvalue problem was solved by a Galerkin procedure wherein the dependent variables are expressed as a summation of trial functions as shown below,

$$U = \sum_{n=1}^N c_n \varphi_n, \quad T = \sum_{n=1}^N d_n \psi_n, \quad p = \sum_{n=1}^N e_n \zeta_n. \quad (10)$$

Similar calculations by Bühler *et al* (1979) using a more complete set of trial functions showed that the critical Rayleigh number is a function of the geometrical aspect ratios. The critical Rayleigh number for the onset of motion decreases with an increase in the aspect ratio and is the lowest for the infinite layer case which is 1707.8. The critical Rayleigh numbers for cylindrical lateral walls was calculated by Charleson & Sani (1970).

#### 5. Transition to time dependence

When the Rayleigh number is increased beyond a certain critical value, RB convection becomes oscillatory through an instability mechanism predicted for the two-dimensional

infinite layer case by Clever & Busse (1974, 1987) using linear stability theory. Stability analysis beyond oscillatory instability has not been reported. However, experimental results of large aspect ratio enclosure flow by Ahlers & Behringer (1978) indicate that the flow becomes turbulent soon after oscillatory convection. For intermediate aspect ratio containers (aspect ratios between 5 and 30), experiments by Oertel (1982, pp. 3–24) indicate a spatially complex structure. Walden *et al* (1984) discovered from their measurements that the temporal behaviour of the fluid was spatially dependent.

In contrast, RB convection in small aspect ratio as reported in experiments (Maurer & Libchaber 1979; Gollub & Benson 1980; Bergé *et al* 1982, pp. 123–148) indicate that the dynamical behaviour of thermal convection follows a well-defined set of bifurcations to chaos and turbulence as the Rayleigh number is increased. The dynamical behaviour is simpler due to the restricting influence of the side-walls. A more interesting general observation was that the bifurcation sequences were qualitatively similar to simple dynamical systems such as nonlinear ordinary difference and differential equations. Examples of bifurcation sequences or scenarios include the Ruelle-Takens scenario (Ruelle & Takens 1971) and the Feigenbaum sequence (Feigenbaum 1978). For a more complete documentation of bifurcation in simple dynamical systems, the reader is referred to Parker & Chua (1988).

A brief description of the experiments is in order. In the experiments of Gollub & Benson (1980), two enclosures  $3.5 \times 2.1 \times 1$ , and  $2.4 \times 1.2 \times 1$ , and two different Prandtl numbers (2.5 and 5) were investigated. Four major transition sequences to turbulent flows were reported. One significant result was that the transition depended not only on the Prandtl number and geometry but on the initial conditions as well. The number of different flow configurations was surprisingly many for a small aspect ratio box.

Experiments on a small box ( $2 \times 1.2 \times 1$ ) with silicone oil ( $Pr = 130$ ) were carried out by Bergé *et al* (1982, pp. 123–48) and Arroyo & Savirón (1992). The observed routes to chaos were similar to those observed for fluids with lower Prandtl numbers such as air (Kirchartz & Oertel 1988). Maurer & Libchaber (1979) performed experiments with liquid helium ( $Pr = 0.5$ ) and observed similar routes to turbulence. Although the dynamical behaviours were not identical, one common feature noted was that the first bifurcation to time-dependent behaviour is oscillatory with a single independent frequency. Furthermore, the critical Rayleigh number for the onset of time-dependent flow decreased with a reduction of the fluid Prandtl number.

## 6. Relevance of numerical studies

The central problem confronting a fluid dynamicist is the phenomena of turbulence. Due to the unacceptably fine mesh required to resolve the length scales of turbulence, it is impossible to numerically simulate turbulence in flows that occur in practical applications, given present computing resources. One approach to understanding the phenomena of turbulence better would be to completely solve a simpler problem that nevertheless shares many of the features of turbulence. One such “simpler problem” is the Rayleigh-Bénard problem in small aspect ratio enclosures in the unsteady and chaotic regime. Nevertheless, computations in an RB system are not particularly easy.

Such computations must necessarily be three-dimensional using numerical schemes with low numerical diffusion and hence higher order and computationally expensive.

Also, since the flow is time-dependent, the integration with respect to time must be carried out for long intervals. With the phenomenal increase in computing power in the past several years, these problems have become tractable and represent a very exciting area of research. By an accurate numerical simulation of well-documented bifurcation experiments, one could begin to understand the phenomena of bifurcation and transition better. Insights into the physics of the RB system could potentially lead to a better understanding to the important problem of transition to turbulent flows.

The study of bifurcation phenomena in small enclosures is also a fascinating problem in its own right. Busse and coworkers have more or less identified, in a complete manner, the instabilities associated with roll convection in infinite enclosures. The extension to finite boxes, especially small boxes, is not straightforward, and practically intractable from the point of view of a stability problem since the base state is analytically complex. The only tool available is a complete and accurate numerical simulation of the basic equations. The documentation of all instability mechanisms associated with small and intermediate enclosures represents a formidable and challenging enterprise for applied mathematicians, physicists and engineers.

## 7. Early numerical study of the RB system

Numerical investigation of the bifurcation phenomena in small boxes has been few and very recent. Upson *et al* (1981, pp. 245–59) used a modified Galerkin finite element method to simulate thermal convection for the case experimentally investigated by Maurer & Libchaber (1979) for a  $3.29 \times 1.8 \times 1$  enclosure and a Prandtl number of 0.5. The frequencies of oscillations were close to that in the experiments. Two different flow patterns, one with two rolls and the other with three rolls were found to occur for the same boundary conditions (but different initial conditions). However, the transition from three to two rolls that was documented experimentally could not be reproduced. Mukutmoni & Yang (1992) numerically simulated the transition from three to two rolls. It is possible that the simulations of Upson *et al* (1981, pp. 245–59) were not carried out long enough for the instability to manifest itself.

Kessler (1987) carried out simulations with a 4:2:1 box for air. Spectral Galerkin methods were used. The dependent variables were expanded in the form similar to (10),

$$\psi = \sum_{i=1}^L \sum_{j=1}^M \sum_{k=1}^N a_{ijk}(t) C_i(x) C_j(y) C_k(z). \quad (11)$$

The  $C(x)$ 's are the so-called 'beam functions' that satisfy the zero normal and tangential velocity at the wall. As given in Chandrasekhar (1961), they have the following functional form:

$$C(x) = (\cosh \lambda_m x / \cosh(\lambda_m/2)) - (\cos \lambda_m x / \cos(\lambda_m/2)) \text{ or } (\sinh \mu_m x / \sinh(\mu_m/2)) - (\sin \mu_m x / \sin(\mu_m/2)). \quad (12)$$

$\lambda_m$  and  $\mu_m$  are the eigenvalues for a specific set of boundary conditions. For higher Rayleigh numbers where extra resolution is required for the boundary layers, Chebyshev polynomials were used. The numerical simulations looked into details of the steady and oscillatory convection. However, *ad-hoc* symmetry conditions were imposed to



reduce computational costs. As a result, it was not a suitable method for studying flow transitions. In particular, the symmetry conditions forced an odd number of rolls.

Yahata (1984) computed for several aspect ratios including cylindrical side walls using a Galerkin method. The assumption of slip walls made the computations more efficient but less realistic. The calculations showed several possible routes to chaotic flow. Non-Boussinesq effects were investigated.

Urata (1986) used spectral methods to compute for a 2:2:1 box for Prandtl numbers of 1 and 3. The study looked into the details of time-dependent convection. However, the assumption of slip-walls meant that the solution obtained was somewhat unrealistic. The computations spanned the chaotic regime. The Lyapunov exponents were calculated. For  $Pr = 1$ , and  $Ra = 20,000$ , the dimension of the chaotic attractor was determined to be 3.3.

## 8. Some recent studies on small aspect ratio RB convection

In this section we review some of the recent works on RB convection in small boxes. All studies reviewed in this section computed realistic cases (non-slip walls). More significantly, extensive comparisons with experiments were made. The RB system even for small aspect ratio enclosures permits multiple solutions to the numerical problem. Although some solutions computed without proper experimental validation are real, many of them are obtained from incorrect application of boundary conditions or inadequate resolution from a coarse mesh and are therefore not physical. If the purpose of the study is to understand the physics, comparison and validation from closely related experiments is imperative. As observed by Yang & Mukutmoni (1992, pp. 23–41), only then can it be claimed that the numerical solution is indeed an actual solution.

Mukutmoni & Yang (1993a) studied a two-roll RB convection for the case experimentally studied by Gollub *et al* (1980, pp. 22–7) for a 3.5:2.1:1 box for a fluid of Prandtl number 2.5. The schematic diagram of the two-roll RB convection is shown in figure 2.

Due to the nonlinearity of the governing equations, the solution is not uniquely determined by the governing parameters. The solution depends on the initial conditions.

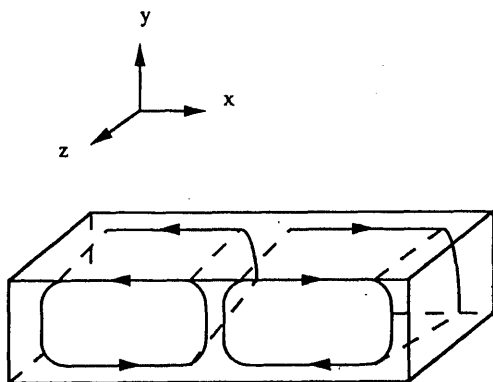
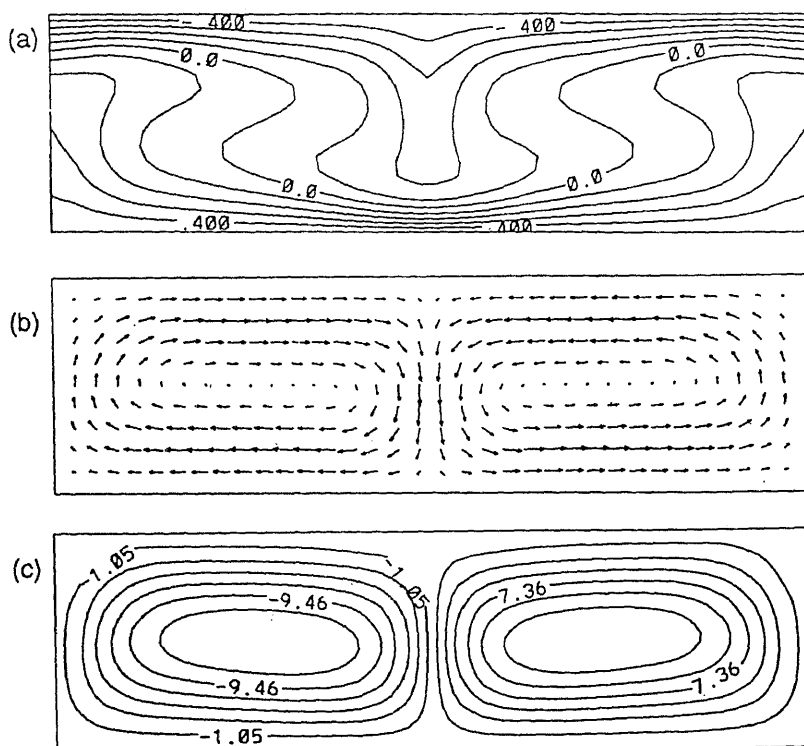


Figure 2. Schematic of a two-roll RB convection.

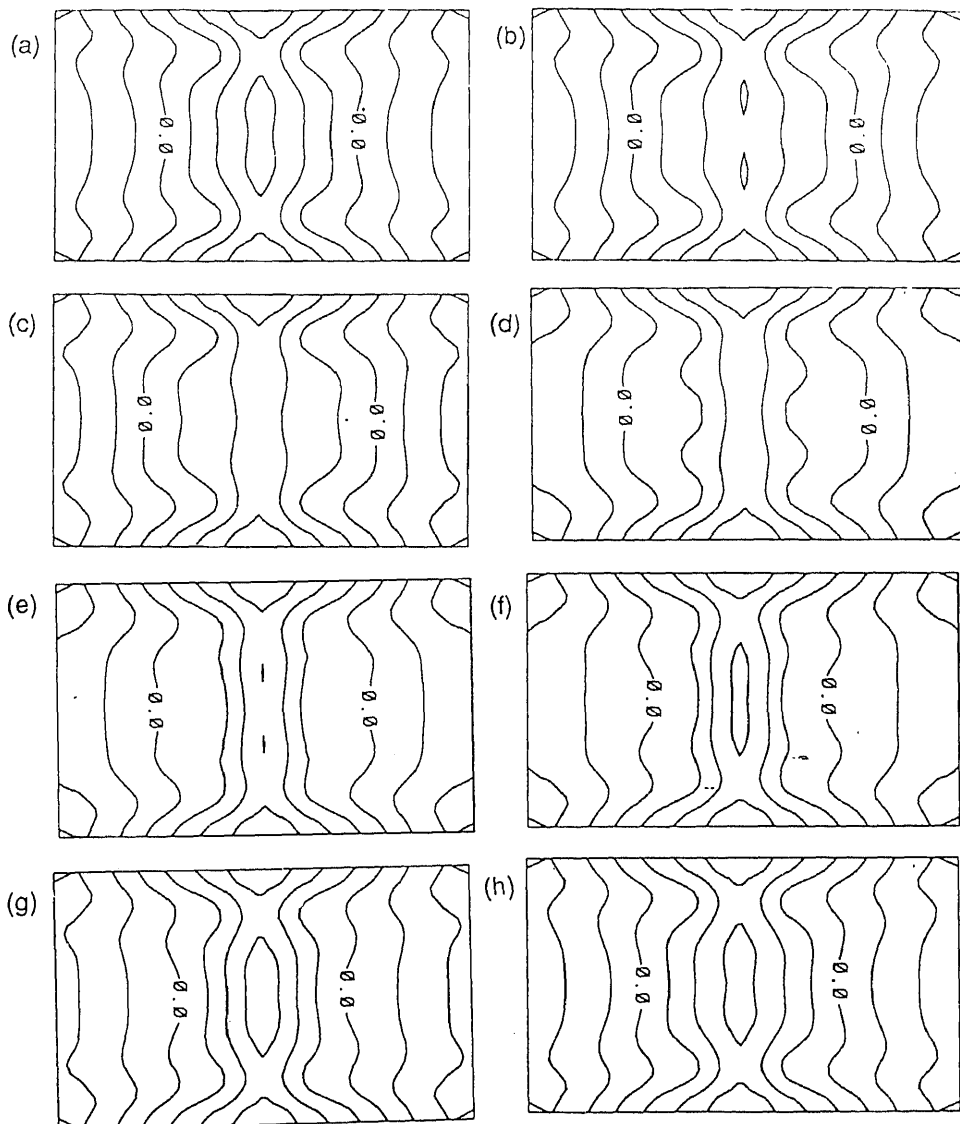


**Figure 3.** (a) Sectional streamlines, (b) velocity vectors, and (c) isotherms at vertical section  $z = 1.7$ ;  $Ra = 20,000$  (Mukutmoni & Yang 1993a).

Using suitable velocity perturbations, a two-roll flow pattern (figure 3) was developed. Based on a grid refinement study (Mukutmoni & Yang 1993a), a  $20 \times 20 \times 20$  grid was used for the computations. A third-order finite-difference scheme QUICK (Leonard 1983, pp. 211–26) was used. The finite-volume method and SIMPLEX algorithm (Van Doormal & Raithby 1984) were implemented. Good quantitative agreement with experiments was found with respect to the frequencies and critical Rayleigh numbers. Qualitatively, the behaviour was identical. Both the computations and experiments showed a Hopf bifurcation followed by a period-doubling transition.

The numerical results showed that the oscillating temperature and velocity field exhibited a standing wave pattern propagating along the axis of the rolls. The nature of the oscillating temperature field is shown along a horizontal section in figure 4 as a series of contour plots spanning one complete oscillation cycle. Another study by Mukutmoni & Yang (1992) revealed a similar standing wave pattern in a  $3.3:1.9:1$  box for a fluid Prandtl number of 0.5 (figure 5). These computations were based on the experiments of Maurer & Libchaber (1979). Linear stability theory (Clever & Busse 1974) predicts a travelling wave pattern for a horizontally unbounded domain. The standing wave observed in the simulations is an obvious extension of their results.

Several experiments of Rayleigh–Bénard convection in small enclosures (Gollub *et al* 1980, pp. 22–7; Libchaber & Maurer 1981, pp. 259–86; Libchaber *et al* 1982) have shown that among the several bifurcation sequences observed there is one that approximates the classic period-doubling route to chaos observed in iterated maps

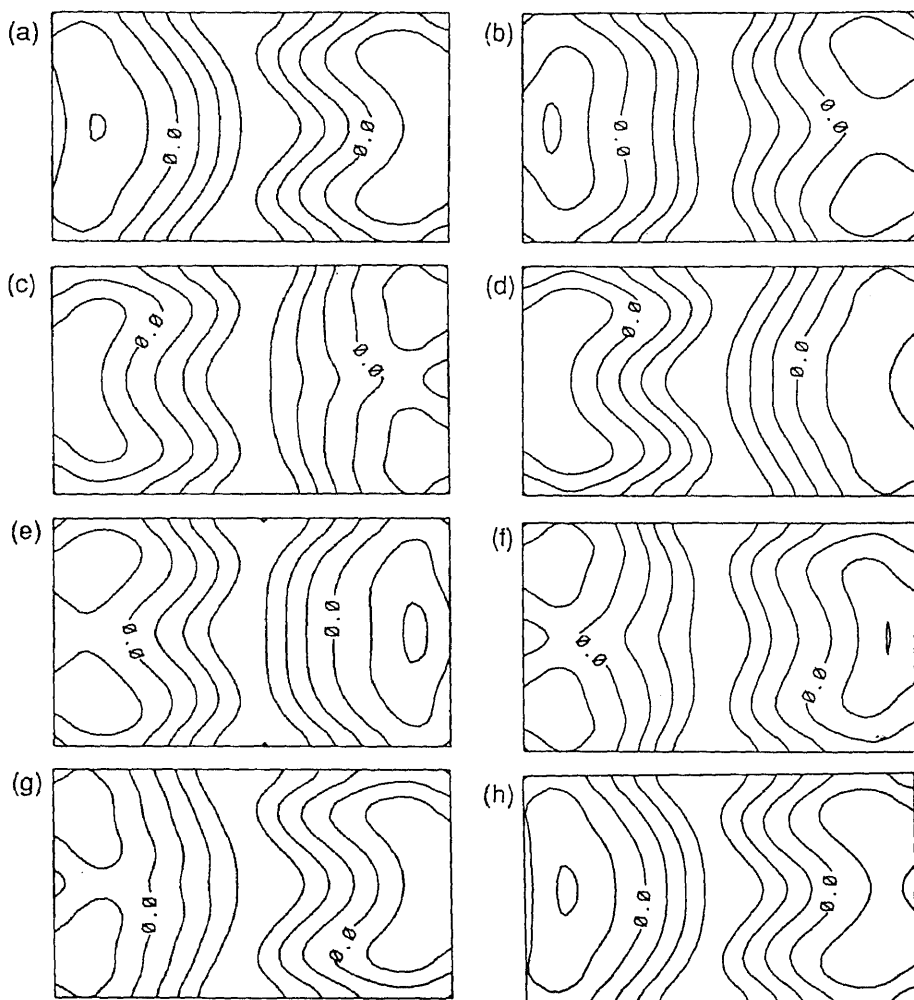


**Figure 4.** Instantaneous isotherms over a complete oscillation cycle;  $Ra = 35,000$  at the horizontal section  $y = 0.78$  (Mukutmoni & Yang 1993a).

by Feigenbaum (1978). As an example, the following quadratic map exhibits such a behaviour as the control parameter  $\lambda$  is increased,

$$x_{n+1} = \lambda x_n(1 - x_n), \quad (13)$$

where  $0 \leq x_n \leq 1$ . In such a route to chaos (known as the Feigenbaum route), one observes an infinite cascade of period-doubling bifurcations. The successive critical control parameters get smaller and are in a geometric progression. A sufficiently larger number of bifurcations. The ratio is known as the Feigenbaum constant.

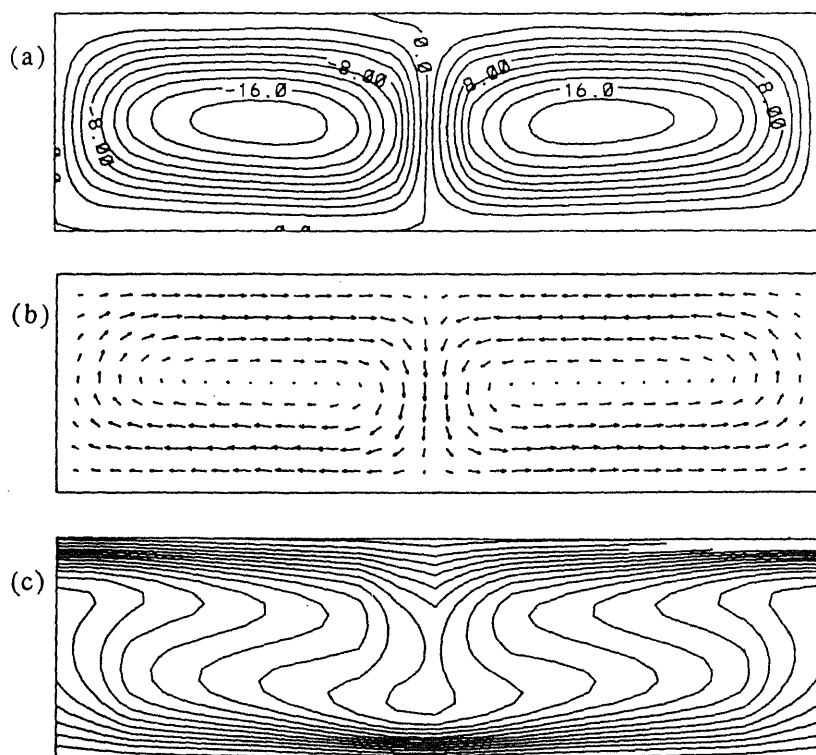


**Figure 5.** The oscillating isotherms over one complete period at the horizontal section  $y = 0.8$ . Time interval is 0.023 (Mukutmoni & Yang 1992)

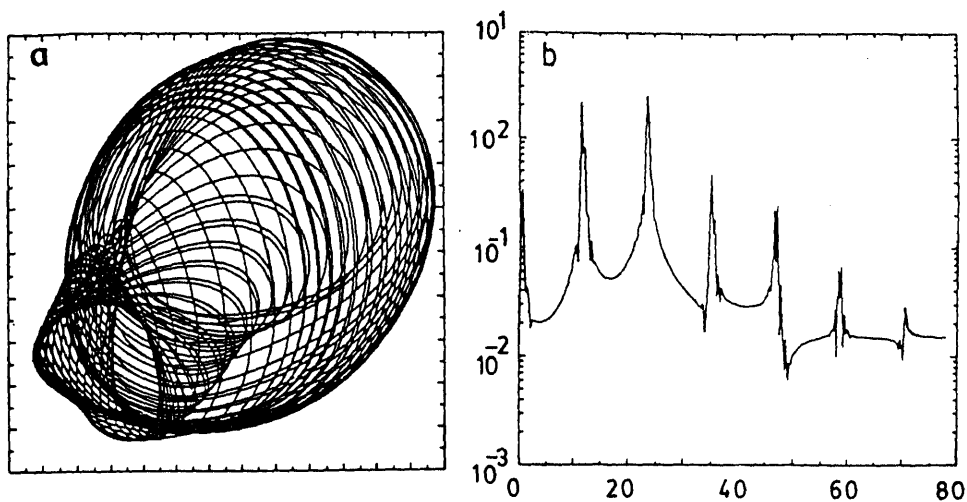
number. Thus,

$$\lim_{n \rightarrow \infty} (\lambda_n - \lambda_{n-1}) / (\lambda_{n+1} - \lambda_n) = 4.669201609 \dots \quad (14)$$

For the Rayleigh-Bénard system the sequence of sub-harmonic bifurcations experimentally observed is not an infinite cascade but stops at not more than five. This is followed by a bifurcation to quasi-periodicity and chaos. The reason why the Feigenbaum sequence is reproduced only partially for the RB system was looked into by us (Mukutmoni & Yang 1993b). In our investigations (Mukutmoni 1991; Mukutmoni & Yang 1993b), it was discovered that the bifurcation sequence strongly depended on the step increase in the Rayleigh number in the numerical study. For a step increase of 2000, only one period-doubling bifurcation was observed (Mukutmoni & Yang



**Figure 6.** Mean velocity and temperature fields for  $Ra = 37,000$ : (a) streamlines (b) velocity vectors, (c) isotherms (Mukutmoni & Yang 1993b).



**Figure 7.**  $Ra = 39,000$ ; (a) phase trajectory, (b) spectral amplitudes (Mukutmoni & Yang 1993b).

1993b). For a smaller step increase, two subharmonic bifurcations were reported (Mukutmoni 1991).

After bifurcation to quasi-periodic flow, it was shown that the two-roll pattern developed asymmetry at  $Ra = 37,000$  between the two rolls (figure 6). It was also reported that the independent frequency added to the system was an order of magnitude smaller than the first fundamental frequency (figure 7). The likely conclusion backed by experiments (Gollub *et al* 1980, pp. 22–7), is that as the Rayleigh number is increased, the symmetry between the rolls cannot be sustained and when that happens the Feigenbaum sequence is terminated. To further test this hypothesis, computation carried out by Mukutmoni & Yang (1993b) artificially imposed such a symmetry. It was found that under such conditions, the Feigenbaum sequence is reproduced. One period-doubling sequence is shown in figure 8. Furthermore, the computations were able to predict the ‘windows’ of periodic flow between chaotic regimes as observed in the quadratic map (Parker & Chua 1988).

It is rather surprising that a complicated system such as RB convection in small boxes for a certain parameter range can dynamically behave similar to one-dimensional nonlinear difference equations. However, as revealed in the experiments (Maurer &

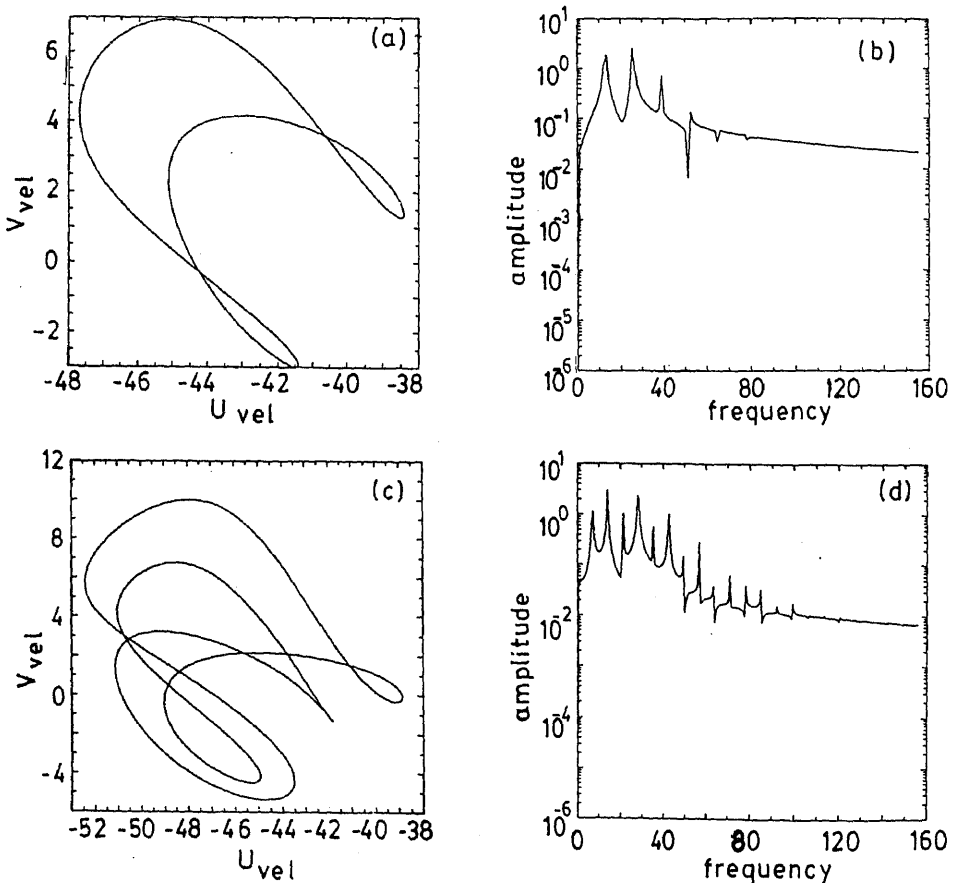
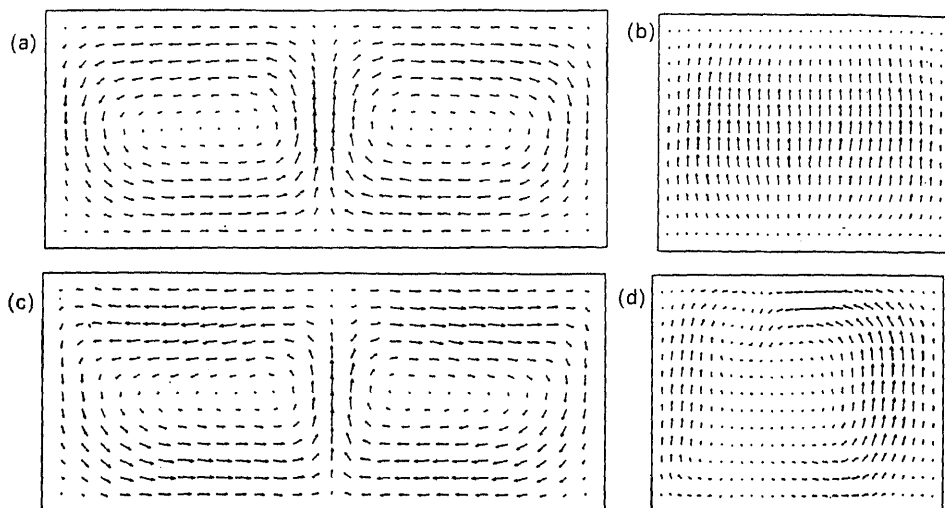


Figure 8. (a) Phase trajectory, (b) spectral amplitudes for  $Ra = 47,000$ ; (c) phase trajectory, (d) spectral amplitudes for  $Ra = 48,000$  (Mukutmoni & Yang 1993b).

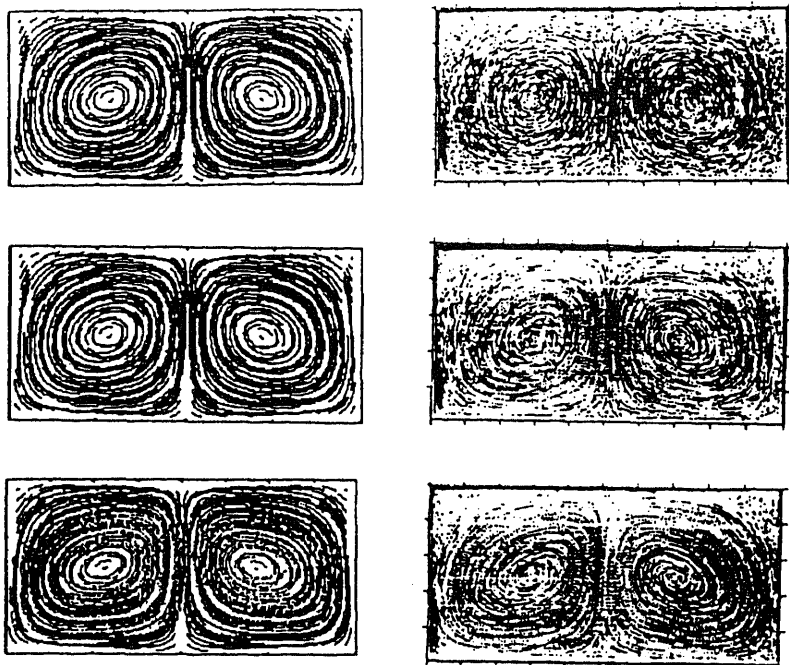
Libchaber 1979; Gollub & Benson 1980; Bergé *et al* 1982, pp. 123–48), the complete dynamical behaviour of the RB system is far richer than simple dynamical systems consisting of ordinary difference and differential equations. This is not surprising, since the equations of thermal convection are governed by a system of coupled nonlinear *partial* differential equations. It is also to be expected that the experiments have not revealed all the possible scenarios. With present computing resources, a thorough investigation of the dynamical behaviour is now tractable, and presents an exciting avenue for research to a fluid dynamicist and to a numerical analyst as a benchmark problem (Mukutmoni & Yang 1991).

Mukutmoni & Yang (1994a) further demonstrated the usefulness of numerical computations to gain specific physical insights when used in conjunction with experiments. Gollub & Benson (1980) uncovered several routes to chaos and turbulence. In all cases, with one exception, the dynamical behaviour increased in temporal complexity as the Rayleigh number was increased. In one apparently anomalous case, for aspect ratios of 2.42 and 1.23, and Prandtl number of 5.0, the flow reverted from a quasi-periodic to steady-state with an increase in Rayleigh number. Mukutmoni & Yang (1994a) showed that the unexpected reversion to steady-state was caused by a bifurcation in the spatial pattern that stabilized the flow. Figure 9 shows the mean velocity and temperature field, before and after the bifurcation.

Preliminary numerical investigation into high Prandtl number fluids using the experiments of Arroyo & Savirón (1992) was attempted by Mukutmoni *et al* (1993). Figure 10 shows the comparison with experiments in terms of pathlines. The experiments and numerical simulations were for an aspect ratio of 2.03 and 1.19 for silicone oil ( $Pr = 130$ ). One numerical difficulty associated with high Prandtl fluids is that time-dependence occurs at higher Rayleigh-numbers and that requires a finer grid and adequate resolution of the boundary layers.



**Figure 9.** Velocity vectors in (a)  $x-y$  plane for section  $z=0.7$ , and (b)  $z-y$  plane for the section  $x=1.0$ ,  $Ra=40,000$ , and in (c)  $x-y$  plane for section  $z=0.7$ , and (d)  $z-y$  plane for the section  $x=1.0$ ,  $Ra=130,000$  (Mukutmoni & Yang 1994a).



**Figure 10.** Computed pathlines (Mukutmoni *et al* 1993) on the left. Experimental results (Arroyo & Savirón 1992) on the right. Rayleigh numbers starting from the top, 6319, 11101 and 22884.

Thus far, the discussion has been mostly restricted to bifurcation in the temporal behaviour of the system. In the following section we discuss ongoing research in the study of the bifurcation in spatial patterns.

## 9. Pattern selection in Rayleigh–Benard system

It has long been known that the RB problem is degenerate, i.e., for the same set of governing parameters many solutions are possible (Busse 1978). This is a consequence of the nonlinearity of the problem. The mechanism on how different solutions to the problem evolve and compete and the process by which a particular flow configuration undergoes changes is broadly known as pattern selection. Some of the earliest work on the pattern selection problem was done by Busse and coworkers (Busse & Whitehead 1974; Clever & Busse 1974; Busse & Clever 1979). In their works, they identified several instability mechanisms that influence the pattern selection process of RB convection in an infinite horizontally unbounded media. However, for an infinite case, the number of stable solutions is infinite. For roll convection, the stable solutions occur over a finite bandwidth of wavenumbers.

For large aspect ratio boxes (aspect ratios greater than 30) it has been experimentally shown (Busse 1978) that the instability mechanisms documented for the horizontally unbounded case in general apply, although the number of solutions is finite but large. For smaller enclosures, the evidence is insufficient. Experimental results indicate that



at least some of the mechanisms do apply (Kolodner *et al* 1986; Kirchartz & Oertel 1988). However, it is to be expected that new instability mechanisms would occur in the presence of lateral walls.

There have been several attempts to solve the problem with useful approximations of the governing equations. Using a perturbation expansion valid for regimes slightly above the critical, Newell & Whitehead (1969) and Segel (1969) derived an amplitude equation. Subsequent researchers have improved and generalized this amplitude equation to study growth and saturation of the rolls. Using this approach, Greenside & Coughran (1984) predicted (experimentally confirmed by Gollub & Heutmaker 1984) that rolls tend to intersect rigid non-slip walls in an approximately perpendicular direction. The amplitude equation is a two-dimensional nonlinear partial differential equation that approximates (1) to (5). It has the following form,

$$\partial\phi/\partial t = [\varepsilon - (\nabla^2 + 1)^2]\phi - \phi^3, \quad (15)$$

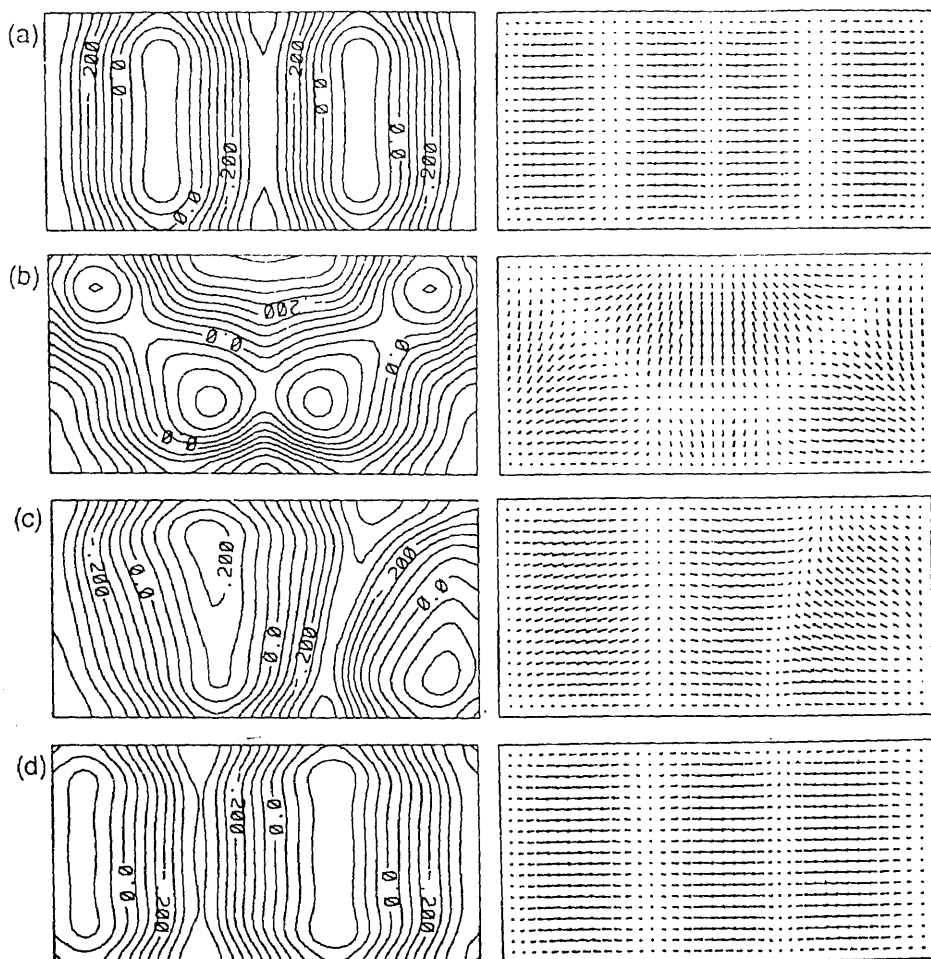
where  $\varepsilon$  is the perturbation parameters  $(Ra/Ra_c) - 1$ . The dependent variable  $\phi$  is the amplitude.

Greenside & Coughran (1984) studied the pattern selection problem by numerically solving the amplitude equation and imposing random initial conditions and integrating in time. The results indicate that irregular patterns or defects do develop but they smoothen out for large Rayleigh numbers. The stationary solutions were found to be symmetric. It was determined that the time scales needed to achieve a stationary solution was proportional to the horizontal diffusion scale.

Nevertheless, the amplitude equations are valid only slightly above the critical Rayleigh number. To really model the phenomena accurately in the high Rayleigh number regime, the full Boussinesq equation must be solved. The problem is less tractable than a small box since a large computational grid is required. Furthermore, since the relaxation time is proportional at least to the horizontal diffusion time, the simulations must be carried out for a longer time.

RB convection in an intermediate aspect ratio box was experimentally investigated by Kolodner *et al* (1986). They concluded that the observations were consistent with the results of two-dimensional stability theory, provided one makes plausible allowances for the experimental imposition of finite lateral boundaries. The focus of most experimentalists has been on roll convection. One well-documented phenomenon experimentally observed in small and intermediate boxes is the decrease in the number of rolls with an increase in Rayleigh number (Kolodner *et al* 1986; Leith 1987; Kirchartz & Oertel 1988). According to linear stability theory, the loss of roll phenomena, as it is sometimes referred to, is an outcome of the skewed-varicose instability (Busse 1978).

The loss of rolls was numerically investigated by Mukutmoni & Yang (1992) for a small box (4:2:1) for a Prandtl number of 0.71. The transition from 4 to 3 rolls was documented as shown in figure 11. The transition sequence shows the typical slanting of the rolls as well as the thinning and thickening of the distorted rolls in the time sequence. In our computations (Mukutmoni & Yang 1994b), using a  $98 \times 20 \times 50$  grid, we simulated a transition from 10 to 8 rolls, for an enclosure of aspect ratios 10.61 and 5.32 and Prandtl number of 3.5. The transition sequence is shown in figure 12. Our results (Mukutmoni & Yang 1994) also showed that there is a generation of vertical vorticity or swirl in the transition process. This means that



**Figure 11.** Transition sequence from 4 to 3 cells.  $Ra = 10,000$ ; horizontal section  $y = 0.8$  (Mukutmoni & Yang 1992).

the transition process is not governed entirely by the skewed-varicose instability for which there is no vertical vorticity (Busse 1978). The numerical computations were an accurate simulation of the experiments of Kolodner *et al* (1986).

Another related issue of roll convection is related to the alignment of the rolls. Early theoretical (Davis 1967) and experimental results (Stork & Müller 1972) seemed to suggest that only rolls parallel to the short side of a rectangular container are stable. However, we (Mukutmoni & Yang 1992) showed that rolls parallel to the long side for a small enclosure of aspect ratio 3.5 and 2.1 and Prandtl number 2.5 are stable below a certain critical Rayleigh number. The conclusion of the study (Mukutmoni & Yang 1992) was that long rolls are metastable (unstable to finite perturbations) and definitely less stable than rolls parallel to the short side. Experiments of Kolodner *et al* (1986) found stable rolls parallel to the long side, for intermediate aspect ratio boxes.

Pattern selection and flow transitions in intermediate aspect ratio boxes are much more complicated than smaller boxes due to the less restrictive influence of the

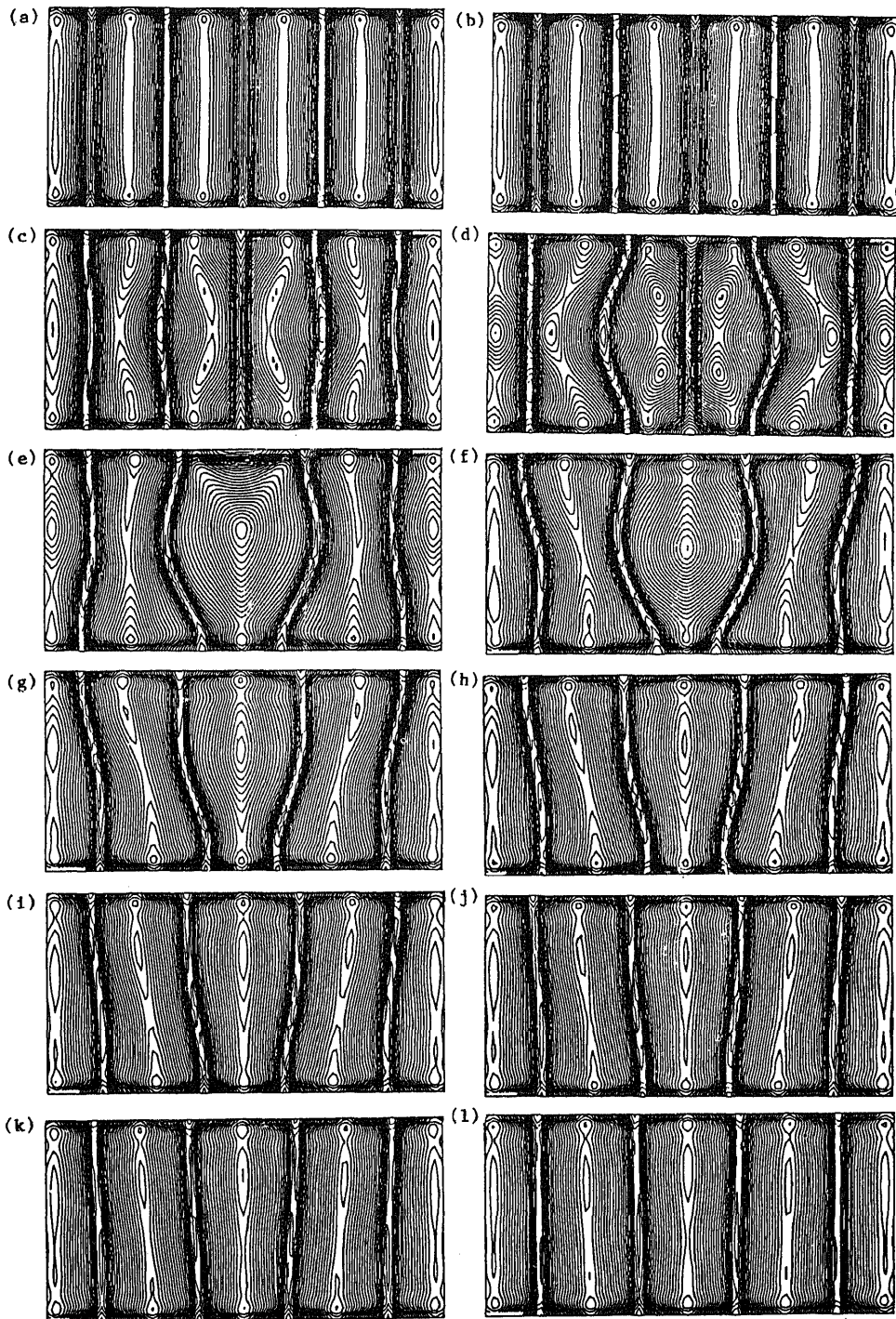


Figure 12. Pattern transition at  $Ra = 17,000$  represented in terms of isotherms at horizontal section  $y = 0.7$ . Time interval of  $t = 1$  (Mukutmoni & Yang 1994b).

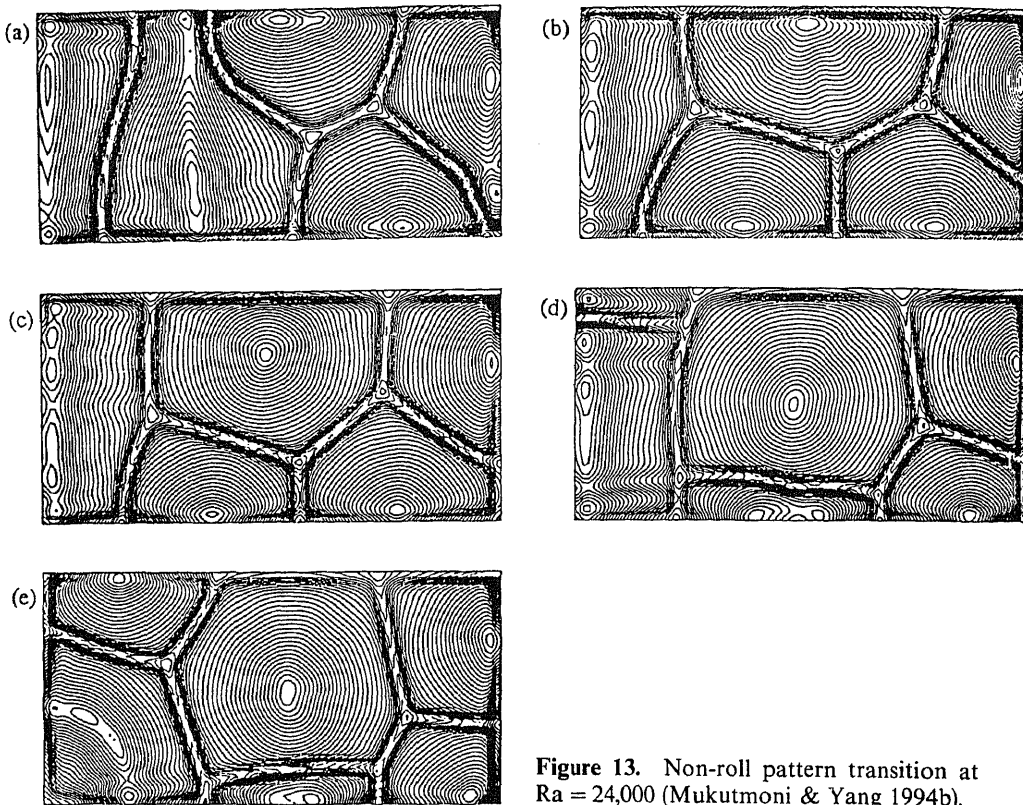


Figure 13. Non-roll pattern transition at  $Ra = 24,000$  (Mukutmoni & Yang 1994b).

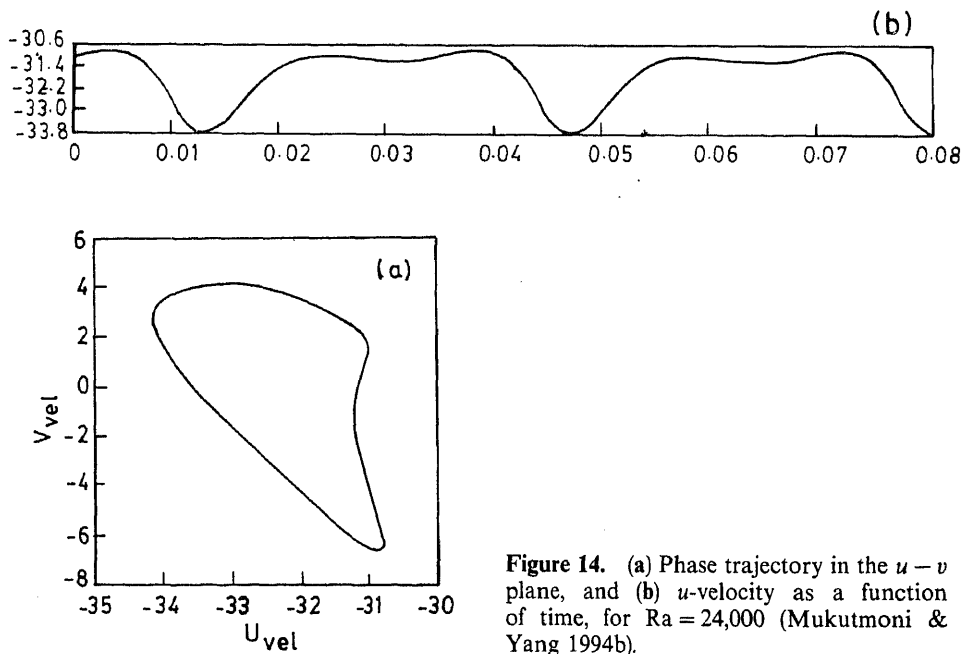


Figure 14. (a) Phase trajectory in the  $u-v$  plane, and (b)  $u$ -velocity as a function of time, for  $Ra = 24,000$  (Mukutmoni & Yang 1994b).

sidewalls. Some experiments hint at a much more complicated temporal behaviour (Walden *et al* 1984) at higher Rayleigh numbers than small enclosures. The flow patterns are more complicated. Several non-roll patterns have been documented. For moderately high Prandtl fluids, bimodal convection has been documented experimentally (Busse 1978). Another interesting non-roll pattern is the "spoke-pattern" convection (Kolodner *et al* 1986) which has a highly complicated three-dimensional structure and is time-dependent as well. We (Mukutmoni & Yang 1994b) documented steady oscillatory convection with a polygonal planform. The planform is shown in figure 13 and the dynamical behaviour is shown in figure 14. Such a pattern has however not been observed in experiments.

## 10. Research needs and concluding remarks

From our discussion it appears that numerical simulation of RB convection in enclosures is a very worthwhile and rewarding research area. However, it is clear that the numerical studies have yet to catch up with the range and scope of the experiments. The review shows that successful simulations have been performed for small aspect ratio enclosures up to the chaotic regime for some cases. Nevertheless, simulations have not covered the entire range of dynamical behaviour. Although experiments are available, there has been no study investigating the intermittency and phase-locking behaviour that has been observed. Systematic and extensive studies of high-Prandtl fluids in enclosures are absent thus far.

It is also worthwhile to look at convection in small enclosures for non-rectangular geometries such as with cylindrical lateral walls or polygonal domains. There has not been any study as yet. Even for rectangular domains, turbulent convection in enclosures at the lower Rayleigh number and (sometimes referred to as "soft" turbulence) has not received the attention it deserves. All numerical studies on turbulent thermal convection use some degree of approximation such as periodicity in the horizontal directions or slip horizontal walls such as the works of Sirovich & Park (1990). One obvious problem for RB convection for high Rayleigh numbers in the turbulent regime is the tractability. However, with the greater availability of powerful workstations and massively parallel machines these problems would be feasible, and lead to a greater understanding of the physics.

For intermediate aspect ratio enclosures, the computational problem is more difficult. Not surprisingly, there have been extremely few studies that tackle the problem in all its complexity. However, due to the more complex physics associated with the problem, it is a more rewarding area. The phenomena of RB convection in intermediate boxes are not well known. The first challenge would be to take a closer look at the experimentally observed flow patterns. In particular, the non-roll patterns have not been looked into at all.

Thermal convection at high Rayleigh numbers is a computationally difficult problem. In this article, we have described, in certain detail, some of the computational techniques used in the various studies. We make no specific recommendations on the type of numerical technique employed. However, higher order methods are to be preferred in order to minimize numerical diffusion. The class of numerical techniques used by researchers include spectral, finite element and finite volume methods. For a complete exposition on the spectral methods the reader is referred to the monograph by Canuto *et al* (1988). An excellent reference for the finite volume method is the book by Fletcher (1988).

Higher order finite difference and finite volume methods have only recently been applied to numerical problems. As an example Lele (1992) describes the compact differencing schemes. Rai & Moin (1991) illustrate higher order upwind-weighted schemes for simulating turbulent flow. The issue of the required mesh density is highly problem dependent. For the RB problem, as described earlier, it depends strongly on the Rayleigh number and the Prandtl numbers. Experimental results are needed to primarily validate the grid resolution.

### List of symbols

$A_x, A_z$	$L_x/L$ and $L_z/L$ , aspect ratios in the $x$ and $z$ directions respectively;
$g$	acceleration due to gravity, $\text{m/s}^2$ ;
$L$	height of the enclosure, $\text{m}$ ;
$L_x, L_z$	dimensions of enclosure in $x$ and $z$ directions ( $\text{m}$ ) respectively;
$\text{Pr}$	Prandtl number;
$\text{Ra}$	Rayleigh number;
$\text{Ra}_c$	critical Rayleigh number;
$t$	non-dimensional time;
$T$	non-dimensional temperature;
$\Delta T$	temperature difference, $\text{K}$ ;
$\mathbf{U}$	non-dimensional velocity vector;
$u, v, w$	non-dimensional $x$ -, $y$ - and $z$ -direction velocities respectively;
$x, y$	non-dimensional horizontal and vertical spatial coordinates respectively;
$z$	non-dimensional spatial coordinate in the direction of depth;
$\alpha$	thermal diffusivity, $\text{m}^2/\text{s}$ ;
$\beta$	coefficient of volume expansion, $1/\text{K}$ ;
$\varepsilon$	perturbation parameter;
$\lambda$	bifurcation parameter;
$\nu$	kinematic viscosity, $\text{m}^2/\text{s}$ ;
$\rho$	density, $\text{kg/m}^3$ ;
$\phi$	non-dimensional amplitude of perturbation.

### References

- Ahlers G, Behringer R P 1978 The Rayleigh-Bénard instability and the evolution of turbulence. *Suppl. Prog. Theor. Phys.* 64: 186-201
- Arroyo M P, Sàvìrón J M 1992 Rayleigh-Bénard convection in a small box: Spatial features and thermal dependence of the velocity field. *J. Fluid Mech.* 235: 325-348
- Bénard H 1900 Les Tourbillons Cellulaires Dans Une Nappe Liquide. *Revue Générale des Sciences Pures et Appliquées* 11: 1261-1271 and 11: 1309-1328
- Bergé P, Dubois M, Croquette V 1982 Approach to Rayleigh-Bénard turbulent convection in different geometries. *Convective transport and instability phenomenon* (eds) J Zierep, H Oertel Jr (Karlsruhe: G Braun)
- Bühler K, Kirchartz K R, Oertel H Jr 1979 Steady convection in a horizontal fluid layer. *Acta Mech.* 31: 155-171
- Busse F H 1978 Nonlinear properties of thermal convection. *Rep. Prog. Phys.* 41: 1929-1967
- Busse F H, Clever R M 1979 Instability of convection rolls in a fluid of moderate Prandtl number. *J. Fluid Mech.* 91: 319-335
- Busse F H, Whitehead J A 1974 Oscillatory and collective instabilities in large Prandtl number convection. *J. Fluid Mech.* 66: 67-79

- Canuto C, Hussaini M Y, Quateroni A, Zang T A 1988 *Spectral methods in fluid dynamics* (New York: Springer-Verlag)
- Catton I 1970 Convection in a closed rectangular region: The onset of motion. *ASME J. Heat Transfer* 92: 186-187
- Chandrasekhar S 1961 *Hydrodynamic and hydromagnetic stability* (Oxford: University Press)
- Charleson G S, Sani R L 1970 Thermo-convective instability in a bounded cylindrical fluid layer. *Int. J. Heat Mass Transfer* 13: 1479-1496
- Clever R M, Busse F H 1974 Transition to time dependent convection. *J. Fluid Mech.* 65: 625-645
- Clever R M, Busse F H 1987 Nonlinear oscillatory convection. *J. Fluid Mech.* 176: 403-417
- Davis S H 1967 Convection in a box: Linear theory. *J. Fluid Mech.* 30: 465-478
- Feigenbaum M J 1978 Quantitative universality for a class of nonlinear transformations. *J. Stat. Phys.* 19: 25-52
- Fletcher C A J 1988 *Computational techniques for fluid dynamics* (New York: Springer-Verlag)
- Gollub J P, Benson S V 1980 Many routes to turbulent convection. *J. Fluid Mech.* 100: 449-470
- Gollub J P, Benson S V, Steinman J F 1980 A subharmonic route to turbulent convection. *Nonlinear dynamics* (ed.) H G Helleman (New York: Academy of Sciences)
- Gollub J P, Heutmaker M S 1984 *Turbulence and chaotic phenomena in fluids* (ed.) T Tatsumi (Amsterdam: North-Holland)
- Greenside H S, Coughran W M Jr 1984 Nonlinear pattern formation near the onset of Rayleigh-Bénard convection. *Phys. Rev.* A30: 398-428
- Kessler R 1987 Nonlinear transition in three-dimensional convection. *J. Fluid Mech.* 174: 357-379
- Kirchartz K R, Oertel H Jr 1988 Three-dimensional thermal cellular convection in rectangular boxes. *J. Fluid Mech.* 192: 249-286
- Kolodner P, Walden R W, Passner A, Surko C M 1986 Rayleigh-Bénard convection in an intermediate-aspect-ratio rectangular container. *J. Fluid Mech.* 163: 195-226
- Leith J R 1987 Successive transitions of steady states in moderate size containers of air heated from below and cooled above. *Bifurcation phenomena in thermal processes and convection* HTD (New York: Am. Soc. Mech. Eng.) vol. 94
- Lele S K 1992 Compact finite difference schemes with spectral-like resolution. *J. Comput. Phys.* 103: 16-42
- Leonard B P 1983 A convectively stable, third-order accurate finite-difference method for steady two-dimensional flow and heat transfer. *Numerical properties and methodologies in heat transfer* (ed.) T M Shih (Washington, DC: Hemisphere)
- Libchaber A, Larouche C, Fauve L 1982 Period-doubling cascade in mercury, a quantitative measurement. *J. Phys. Lett.* 43: L211-L216
- Libchaber A, Maurer J 1981 A Rayleigh-Bénard experiment: Helium in a small box. *Nonlinear phenomena at phase transitions and instabilities* (ed.) T Riste (New York: Plenum)
- Maurer J, Libchaber A 1979 Rayleigh-Bénard experiment in liquid helium: Frequency locking and the onset of turbulence. *J. Phys. Lett.* 40: L419-L423
- Mukutmoni D 1991 *Transitions and bifurcations in Rayleigh-Bénard convection in a small aspect ratio*. PhD dissertation, Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN
- Mukutmoni D, Kelleher M D, Joshi Y K 1993 AMPHIB: A users manual. Technical Report NPS-ME-93-001, Naval Postgraduate School, Monterey, CA
- Mukutmoni D, Yang K T 1991 *Transition to oscillatory flow in Rayleigh-Bénard convection in a three-dimensional box*. ASME Paper No. 91-HT-10
- Mukutmoni D, Yang K T 1992 Wavenumber selection for Rayleigh-Bénard convection in a small aspect ratio box. *Int. J. Heat Mass Transfer* 35: 2145-2159
- Mukutmoni D, Yang K T 1993a Rayleigh-Bénard convection in a small aspect ratio enclosure: Part I-Bifurcation to oscillatory convection. *J. Heat Transfer* 115: 360-366
- Mukutmoni D, Yang K T 1993b Rayleigh-Bénard convection in a small aspect ratio enclosure: Part II-Bifurcation to chaos. *J. Heat Transfer* 115: 367-376
- Mukutmoni D, Yang K T 1994a Thermal convection in small enclosures: An atypical bifurcation sequence. *Int. J. Heat Mass Transfer* (accepted for publication)
- Mukutmoni D, Yang K T 1994b Pattern selection for Rayleigh-Bénard convection in intermediate aspect ratio boxes. *Numer. Heat Transfer* (Accepted for publication)

- Newell A C, Whitehead J A 1969 Finite bandwidth, finite amplitude convection. *J. Fluid Mech.* 38: 279–303
- Oertel H Jr 1982 Thermal instabilities. *Convective transport and instability phenomena* (eds) J Zierep, H Oertel Jr (Karlsruhe: G Braun)
- Parker T S, Chua L O 1988 *Practical numerical algorithms for chaotic systems* (New York: Springer-Verlag)
- Rai M M, Moin P 1991 Direct simulations of turbulent flow using finite-difference schemes. *J. Comput. Phys.* 96: 15–53
- Rayleigh L 1916 Convective currents in a horizontal layer of fluid when the higher temperature is on the under side. *Philos. Mag.* 32: 529–575
- Ruelle D, Takens F 1971 On the nature of turbulence. *Commun. Math. Phys.* 20: 167–192
- Segel L A 1969 Distant side-walls cause slow amplitude modulation of cellular convection. *J. Fluid Mech.* 38: 203–224
- Sirovich L, Park H 1990 Turbulent thermal convection in a finite domain. *Phys. Fluids* A9: 1649–1668
- Stork K, Müller U 1972 Convection in boxes: Experiments. *J. Fluid Mech.* 54: 599–611
- Upton C D, Gresho P M, Sani R L, Chan S T, Lee R L 1981 A thermal convection simulation in three dimensions by a modified finite element method. *Numerical properties and methodologies in heat transfer* (ed.) T M Shih (Washington, DC: Hemisphere)
- Urata K 1986 Low dimensional chaos in Boussinesq convection. *Fluid Dynamics Res.* 1: 257–282
- Van Doormal J P, Raithby G D 1984 Enhancements of the simple method for predicting incompressible fluid flows. *Numer. Heat Transfer* 17: 147–163
- Walden R W, Kolodner P, Passner A, Surko C M 1984 Non-chaotic Rayleigh–Bénard convection with four and five incommensurate frequencies. *Phys. Rev. Lett.* 53: 242–245
- Yahata H 1984 Onset of chaos in the Rayleigh–Bénard convection. *Prog. Theor. Phys. Suppl.* 79: 25–74
- Yang K T 1988 Transitions and bifurcations in laminar buoyant flows in confined enclosures. *J. Heat Transfer* 110: 1191–1204
- Yang K T, Mukutmoni D 1992 Computational aspects of studies of buoyant enclosure flows. *Computers and computing in heat transfer science and engineering* (eds) W Nakayama, K T Yang (Boca Raton, FL: Begell House-CRC Press)



## **Buoyant plane plumes from heated horizontal confined wires and cylinders**

GUY LAURIAT<sup>1</sup> and GILLES DESRAYAUD<sup>2</sup>

<sup>1</sup>Laboratoire de Thermique, CNAM, 292 rue Saint-Martin, 75141 Paris, Cedex 03, France

<sup>2</sup>INSSET, Université de Picardie, 48 rue Raspail, BP 422, 02109 Saint-Quentin, Cedex, France

**Abstract.** Two-dimensional computations are reported for time-dependent laminar buoyancy-induced flows above a horizontal heated source immersed in an air-filled vessel. Two kinds of heated source were considered: a line heat source, modelled as a heat source term in the energy equation, and a heat-flux cylinder of small diameter. First, comparisons are presented for the results obtained for these two heated sources. Rather large discrepancies between the velocity fields appeared in the conduction regime due to the weak plume motion, while close agreements were found in the boundary layer regime. Nevertheless, same types of bifurcations occur with almost identical frequencies, whatever the Rayleigh number. It is concluded that for dimensions of the enclosures, which largely compared with the cylinder radius, the heat source term model is a promising way to study the behaviour of unsteady plumes owing to its simplicity, flexibility, and low computational costs. Second, transitions to unsteady flows were studied through direct flow simulations for various depths of immersion of a line heat source in the central vertical plane of a vessel. Different routes to chaos were shown to occur according to the aspect ratio of the vessel and the depth of immersion of the line source. Three distinct regimes were detected with different underlying physical mechanisms called natural swaying motion, penetrative convection and Rayleigh–Benard-like convection. The first bifurcations associated with these regimes are supercritical Hopf bifurcation, pitchfork bifurcation and subcritical Hopf bifurcation. Comparisons with experimental results of confined buoyant plumes above heated wires show very good agreement with laminar frequency correlations.

**Keywords.** Buoyant plane plumes; line heat source; heat-flux cylinder; numerical plume simulations; route to chaos.

### **1. Introduction**

Free laminar convection from a horizontal line heat source has received increasing attention during past decades due to the importance of this process, which occurs in

many engineering and natural systems of practical interest such as electronic packages, geophysical systems and nuclear reactors. Although a number of analytical and experimental investigations have been devoted to freely rising plumes, many experimental works have been conducted in confined spaces either to minimize environmental effects or to study the behaviour of confined plumes. On the other hand, a literature survey reveals that very little numerical work has been devoted to confined buoyant plumes; the increase in complexity caused by the strong interaction between the plume and its surroundings due to confinement and thermal boundary conditions, in addition to its oscillatory behaviour makes it difficult to perform a parametric study of the different routes to chaos.

Classical self-similar solutions were widely used in early theoretical studies (Fujii 1963; Gebhart *et al* 1970; Fujii *et al* 1973; see also Gebhart *et al* 1988, chap. 3). However, the coupled differential equations for laminar plane plumes were analytically solved in closed form only for specific values of the Prandtl numbers, i.e.  $Pr = 5/9$  (Crane 1959) and  $Pr = 2$  (Fujii 1963; Gebhart *et al* 1970; Fujii *et al* 1973). The self-similar analyses were extended to second-order (Hieber & Nash 1975) and third-order boundary layer theories (Mörwald *et al* 1986). Recently, a correlation for the centreline temperature has been developed by Lin & Cheng (1992). This correlation equation is valid over the entire buoyancy regime and for any Prandtl number.

During the same period, many experiments were conducted in air, water, silicone fluid and spindle oil (Brodowicz & Kierkus 1966; Lyakhov 1970; Schorr & Gebhart 1970; Gebhart *et al* 1970; Fujii *et al* 1973; Nawoj & Hickman 1977). To explain the discrepancies between analytical and experimental results for centreline temperature and velocity distributions, the concept of virtual line source was introduced by Forstrom & Sparrow (1967), Lyakhov (1970), Hieber & Nash (1975) and Yosinobu *et al* (1979). In this way, it was attempted to explain the divergence of the plume flow from the ideal due to the finite diameter of the wire used to simulate the line source. However, this approach raised some controversies (Schorr & Gebhart 1970; Fujii *et al* 1973; Nawoj & Hickman 1977) and did not account for the entire 15–20% difference between experimental and theoretical results; neither end-conduction effects nor decrease of the plume velocity near both ends of the line source could account for them. Yosinobu *et al* (1979) attributed these differences to the heat losses below the wire caused by fluid entrainment. Lyakhov (1970) found only a weak difference on bounding the space below the line source with an impermeable insulating plate; this is more consistent with the boundary layer theory, which did not take into account convective motion below the source. Mörwald *et al* (1986) drew attention to the fact that such discrepancies did not appear from measurements in water. They argued that the Grashof numbers based on diameter and heat rate input of the heat sources were much larger than one for water only, while  $Gr$  were of the order of one for air, spindle oil and silicone fluids. This could be a possible explanation regarding the origin of the discrepancies, since at low Grashof number, the flow field around the wire contradicts the basic assumptions of the boundary-layer theory, i.e. slender plumes.

Forstrom & Sparrow (1967) were the first to observe a naturally swaying motion of a plume which was studied a short time later by Schorr & Gebhart (1970) through flow visualizations. Following the observations of regular swaying plumes, linear stability analysis of freely rising plumes based on the quasi-parallel theory was performed by Pera & Gebhart (1971) and later by Wakitani & Yosinobu (1984), but they failed to find a critical Grashof number. Haaland & Sparrow (1973) and Hieber &

Nash (1975) obtained lower branches of neutral curves (and then critical Grashof numbers) by taking some non-parallel and higher-order effects of the base flow into account in the linear stability analysis. This indicates that two-dimensional disturbances are amplified selectively (see also Gebhart *et al* 1988, chap. 11). More recently, Wakitani (1985) using a non-parallel theory (the WKB method) confirmed their results except for the amplification rate of disturbances within unstable regions. The experimental results of Bill & Gebhart (1975) for a plane plume subjected to naturally occurring disturbances and those of Pera & Gebhart (1971), Yosinobu *et al* (1979) and Wakitani & Yosinobu (1984) for controlled disturbances confirmed the prediction of the linear stability analysis over a wide range of Rayleigh numbers. It has been demonstrated that sufficiently high frequency disturbances are stable, as they are convected downstream.

The transition from a laminar to a turbulent state in a freely rising plume was experimentally investigated by Forstrom & Sparrow (1967), Bill & Gebhart (1975), Yosinobu *et al* (1979) and Neto (1989), and in stable-thermally stratified fluid inside a large enclosure by Noto *et al* (1982). Forstrom & Sparrow (1967) and Bill & Gebhart (1975) determined the beginning of the transition to be at Rayleigh numbers  $Ra_x \cong 3.6 \times 10^8$  and  $Ra_x \cong 8 \times 10^8$  respectively, while for a thermal plume in air Noto (1989) obtained  $7 \times 10^7 \leq Ra_x \leq 1.4 \times 10^8$  from spectral analysis of thermocouple signals and change of the slope of the midplane temperature (the Rayleigh number being based on the vertical distance along the plume, and on the heat rate input). Noto (1989) related the large discrepancies found in the critical Rayleigh number to the different methods used in its determination.

By a different formulation of the onset of instability in plume flow, the buckling theory, Kimura & Bejan (1983) and Yang (1992) demonstrated that at the laminar-turbulent transition the buoyant plume assumes a sinusoidal shape which is characterized by a non-axisymmetric deformation; the sinuous mode having the highest growth rate. Their theoretical arguments are strongly supported by experimental evidence (Kimura & Bejan 1983).

Igarashi & Kada (1977) carried out experiments in confined spaces to investigate the natural convective oscillatory motion of air caused by a heated wire placed concentrically along the axis of a horizontal isothermal cylinder. Different diameters of wires (from 0.29 mm to 1.0 mm) and outer cylinders (from 100 mm to 194.5 mm) were used. The same results were obtained regardless of the diameter of the wire. By performing a dimensional analysis to determine the relationship between the frequency of the oscillatory flow, the thermal and geometrical conditions, and the fluid properties, they concluded that the frequency only depends on the diameter of the outer cylinder and on the heat input, and not on the diameter of the wire or its temperature. Thus, the wall temperature of the outer cylinder was chosen as the reference temperature for the fluid properties rather than the film temperature. Furthermore, a modified Rayleigh number based on the heat flux input and on the radius of the outer cylinder was introduced. The relationship between the non-dimensional frequency and the modified Rayleigh number was found to be  $f = C Ra^{0.4}$ . Igarashi & Kada (1977) also determined the initiation conditions at the onset of oscillations. For a given outer cylinder diameter, they found that the swaying motion did not appear when the heat flux input exceeded some critical values but rather occurred when the temperature difference between the wire and the cooled cylinder exceeded a critical threshold. They suggested that the onset of oscillatory motion is governed by the fluid state close to the wire, whereas the frequency of the oscillations is governed by

the outer boundary conditions far from the wire. Their work was extended by Igarashi (1978) in the case of a line heat source placed concentrically along the axis of a horizontal rectangular chamber. The same conclusions as for the cylinder configuration were arrived at and a correlation having the same power value was established. However, due to the additional geometrical parameter, i.e. the height of the cross-section of the chamber, the  $C$ -coefficient has to be a function of the aspect ratio of the chamber if the Rayleigh number is based on the width. Three distinct values of  $C$  were found corresponding to the three distinct steady flow patterns, giving way to different frequency regimes. Igarashi (1978) also demonstrated the existence of a relationship between the frequency and the middle plane vertical velocity far above the wire and suggested that the plume oscillations are related to the flow circulation rate.

Eichhorn & Vedhanayagam (1982) determined analytically a power value of 0.3 which correlated their experimental results for a water turbulent plume within  $\pm 10\%$ . A  $1/3$  power value was also found both by Urakawa *et al* (1983) for the swaying frequency of a spindle oil plume with a free surface and by Wakitani & Yosinobu (1984) for a laminar air plume. Noto & Matsumoto (1986) and Noto (1989) found a swaying frequency of the thermal air plume proportional to the 0.4 or  $1/3$  power of the heat rate depending on whether the plume reaches the ceiling of the enclosure in a laminar or a turbulent state respectively.

A meandering motion was also noted in experiments on confined plumes by Fujii *et al* (1973, 1982), Eichhorn & Vedhanayagan (1982) and Urakawa *et al* (1983); experiments performed in large parallelepipedic enclosures not only showed that the plume sways in a plane perpendicular to the wire but that it can also meander in the direction of the wire, i.e. across the span of the plume. However it should be noted that the underlying physical mechanisms are not yet clarified. Eichhorn *et al* (1974) and Incropera & Yaghoubi (1980) in experimental studies of immersed isothermal cylinders also observed "transition from 2 to 3-dimensional instability, with increased axial twisting or billowing". This transition was attributed to fluid entrainment effects by Incropera & Yaghoubi (1980). When a meandering motion exists, these two periodic motions (meandering and swaying) are not independent of each other: the swaying motion is stable only when the meandering waveforms along the heater are stable. This happens when the heater length equals integral multiples of a half wavelength, the meandering wave being a fairly precise sine curve (Eichhorn & Vedhanayagam 1982; Urakawa *et al* 1983). The liquid surface height (depth of immersion), linearly related to the meandering wavelength, and the length of the heater are the main parameters controlling the meandering waveform. Pera & Gebhart (1971), Nawoj & Hickman (1977), Yosinobu *et al* (1979), Noto *et al* (1982) and Noto (1989) did not find any meandering motion in their experiments. Noto related the meandering motion to the width of the enclosure apparently without experimental proof: meandering waveforms would appear only for small widths of enclosures. It is worth noting that meandering motions were mainly observed in free-surface fluid experiments with heated wires or cylinders. Fujii *et al* (1973) carried out experiments in air, water and spindle oil and meandering waves only occurred in the case of liquids. Only Bill & Gebhart (1975) mentioned oscillations along the wire span for air plume by interferogram visualizations of plume cross-section.

Numerical and experimental work for heated wires located just below a free surface was recently conducted by Maquet *et al* (1992) and Rozé *et al* (1993). Maquet *et al* (1992) carried out numerical experiments for a square pool with a free surface and

differentially heated sidewalls in which a wire of constant temperature was located below the horizontal free surface. The buoyancy and surface tension mechanisms were incorporated into their formulation. From steady state calculations, they concluded that, even if the temperature differences are small, the free surface deformation is not small with respect to the depth of immersion of the wire. For a 20  $\mu\text{m}$ -diameter wire, Rozé *et al* (1993) experimentally observed a steady surface deformation taking the shape of a trough when the wire was very close to the free surface and the shape of a crest when the depth of immersion was of the order of 1 mm. These two effects were related to surface tension and buoyancy mechanisms respectively. By increasing the wire temperature, an oscillatory motion around the wire arose simultaneously with free-surface propagating waves, these two phenomena having the same frequency. While this bifurcation was recognized as supercritical Hopf bifurcation, they remarked that some liquids (such as silicone oil) developed such oscillatory motions while others (such as water) were not. Moreover, secondary instabilities were not observed beyond the Hopf bifurcation.

Fujii *et al* (1973) were the first to observe a deflection of the plume towards one of the walls when the temperatures of the vertical walls were slightly different. This deflection inhibits the swaying motion. Later on, some experimental studies were devoted to plume interactions (Pera & Gebhart 1975; Incropera & Yaghoubi 1980) or to the influence of solid or liquid interfaces on heat transfer of buoyant plumes (Reimann 1974). Jaluria (1982) investigated the interaction between a plume and a vertical unheated surface and showed that the basic mechanism of the deflection process is the limitation of the flow which supplies the fluid entrained downstream by the plume. Incropera & Yaghoubi (1980) observed various modes of plume interactions from an array of horizontal cylinders: interaction might occur at the air-water surface, before reaching the interface by forming well-ordered ascending and descending flows, or resulting in a highly disordered flow with many recirculating regions of varying sizes. However, the type of plume interaction was shown not to have any influence on the general nature of the temperature distribution.

Although considerable analytical and numerical efforts have been devoted to the study of self-similar solutions for freely-rising plumes, the thermal plumes were confined inside vessels in many experimental investigations. In these cases, the ascending fluid is cooled at a horizontal solid or free surface, inducing a recirculating flow along the sidewalls and an entrainment of underlying fluid. Despite the contributions of the above quoted experimental studies, much remains to be learnt concerning the interaction of the plume motion with its surroundings, in particular with the top surface. To the authors' knowledge, only a few attempts have been made to numerically simulate thermal plumes inside rectangular vessels. Amongst them are the recent studies of Peyret (1990) for double diffusive convection, Xia *et al* (1990, 1991) for an externally heated enclosure containing a local heat source of finite size, and Maquet *et al* (1992) who took into account Marangoni effects. Although most plumes occurring in the environment are turbulent, it is felt that deeper investigation of the behaviour of laminar plumes is justified; since many numerical codes developed to study laminar flows are now used for direct simulations of chaotic or weakly turbulent flows, it is important to ascertain how these laminar models behave. Of course, direct simulations are very time-consuming, even when running on powerful vector computers. Therefore, simplifications of the complexity in the governing equations or in the numerical procedure are not only helpful but necessary. Moreover, the restriction to two-dimensional flow simulations precludes the study of three-

dimensional effects along the line source. Such a limitation is imposed by the computers presently available. However, a two-dimensional model is of interest to provide insight into the occurrence of the swaying motion and also into the transition from periodicity to chaos in considerable detail. Indeed, almost all the experiments highlight a swaying motion of the plume in the cross-section of the wire while very few detect a meandering motion in the spanwise direction of the wire.

Our ultimate purpose being to develop a reliable numerical scheme which would be fast enough to investigate laminar-turbulent transition plume flows and 3D parametric study, it is shown in the present paper that the wire may be modelled as a local source term in the energy equation. Other strategies have been employed such as boundary-fitted coordinates (Himasekar & Bau 1988) or the finite elements method for cylindrical configurations, but at the expense of additional complexities in the numerical treatment of the governing equations. Hybrid coordinate systems have also been used in the vicinity of the cylinder (Fujii *et al* 1982; Farouk & Shayer 1985). Such a procedure allows better control of grid locations but has the disadvantage of the presence of an overlapping zone between the two different grid systems and requires the introduction of somewhat arbitrary conditions where the meshes meet. Another advantage of the local heat source term strategy is its inherent flexibility since it permits the heat source to be moved easily. Also, thermal interactions between two or more line sources inside vessels could readily be studied.

The formulations and the numerical methods used in the present study of dynamical and thermal behaviour of a plume above a heat-flux wire immersed in a rectangular vessel are described in the two following sections (§§2 and 3). The first part of §4 deals with comparisons between results obtained when modelling the wire as a source term in the energy equation or when considering it as a uniform heat-flux cylinder of small diameter. The second part focusses on the different routes to chaos found in square and rectangular vessels according to the depth of immersion of a wire. Almost all these results were presented in four published papers (Lauriat & Desrayaud 1990; Desrayaud & Lauriat 1991, pp. 609–21, 1993; Deschamps & Desrayaud 1994).

## 2. Mathematical formulation

### 2.1 Buoyant plume around a heated cylinder

Consider a two-dimensional fluid-filled vessel of width  $L$  and depth  $H$  enclosed by adiabatic vertical walls and isothermally cooled horizontal surfaces at  $T_c$ . A cylinder of diameter  $d'$  centred at the point  $(x'_s, y'_s)$  generating a heat flux  $Q$  per unit length (W/m) is immersed in the central vertical plane of the vessel (figure 1a). The  $y'$ -axis points upwards. The third dimension of the vessel is taken to be sufficiently large so that a two-dimensional approximation of the flow could be assumed valid.

For a Boussinesq fluid, the conservation equations for mass, momentum and energy are reduced in the dimensionless form to

$$\nabla \cdot \mathbf{V} = 0, \quad (1)$$

$$\frac{1}{\text{Pr}} \left[ \frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V} \cdot \nabla) \mathbf{V} \right] = -\nabla p + \nabla^2 \mathbf{V} + \text{Ra} \theta \mathbf{k}, \quad (2)$$

$$\frac{\partial \theta}{\partial t} + (\mathbf{V} \cdot \nabla) \theta = \nabla^2 \theta, \quad (3)$$

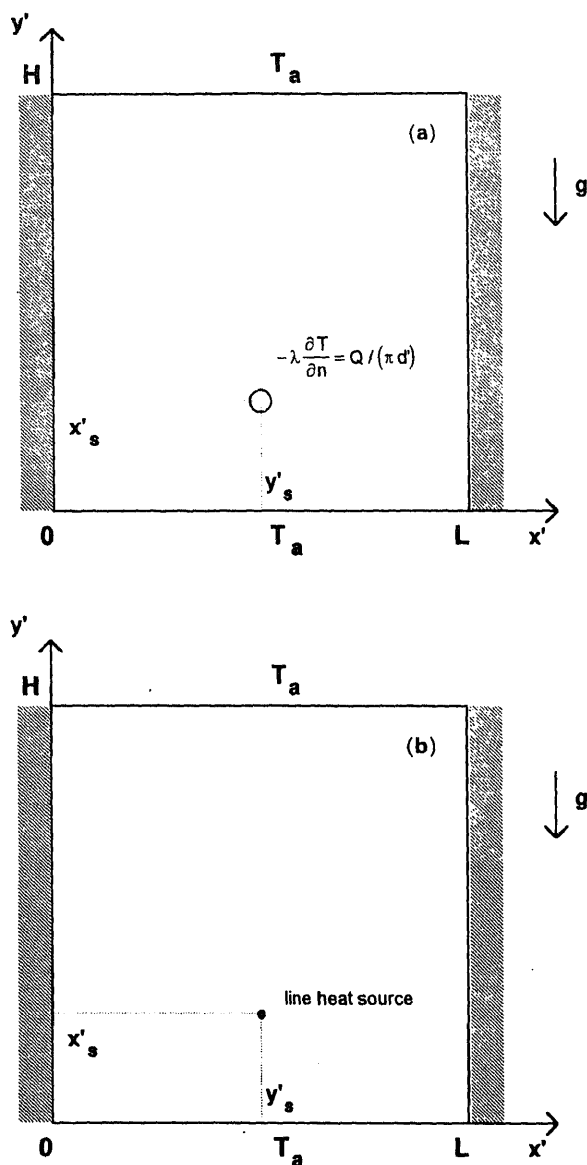


Figure 1. (a) Coordinate system for a cylinder,  $A = 1$ ,  $H_s = 0.75$ ; (b) coordinate system for a line heat source,  $A = 1$ ,  $H_s = 0.75$ .

where the components of the velocity vector  $\mathbf{V}$  are  $(U, V)$  in the  $(x, y)$  directions. The Prandtl and Rayleigh numbers are defined as  $Pr = \nu_0/a_0$  and  $Ra = g\beta QL^3/\lambda_0\nu_0a_0$  respectively.

The dimensionless form of (1)–(3) has been obtained by scaling lengths, time and temperature difference  $(T - T_a)$  by  $L$ ,  $L^2/a_0$  and  $Q/\lambda_0$  respectively. Here  $a_0$  and  $\nu_0$  are the thermal and viscous diffusivities respectively,  $\lambda_0$  the thermal conductivity,  $\rho_0$  the fluid density,  $\beta$  the coefficient of volume expansion and  $g$  the gravitational acceleration; the subscript 0 denotes thermophysical properties at the ambient temperature  $T_a$ , chosen as the reference temperature.

As a result, the relevant hydrodynamic and thermal boundary conditions can be

written in the following dimensionless form

$$\mathbf{V} = 0, \quad \text{at all solid boundaries,} \quad (4a)$$

$$\frac{\partial \theta}{\partial x} = 0, \quad \text{at } x = 0, 1, \quad (4b)$$

$$\theta = 0, \quad \text{at } y = 0, A, \quad (4c)$$

where  $A = H/L$  is the aspect ratio of the cross section of the vessel.

$$\frac{\partial \theta}{\partial \mathbf{n}} = -1/(\pi d), \quad \text{at the cylinder boundary,} \quad (4d)$$

$\mathbf{n}$  being the dimensionless outward unit vector normal to the cylinder surface and  $d$  the dimensionless diameter of the cylinder.

## 2.2 Buoyant plume around line heat source

We now consider two-dimensional convection induced by a line heat source immersed at the dimensionless point  $(x_s, y_s)$  in a Boussinesq fluid. The line source has been modelled as a local source term in the energy equation. This requires that the diameter of a real heat source be much smaller than the dimensions of the vessel. It is convenient to introduce the dimensionless stream function  $\psi$  and the vorticity  $\Omega$  such that

$$\mathbf{V} = (U, V) = (-\partial\psi/\partial y, \partial\psi/\partial x) \quad \text{and} \quad \Omega = -\nabla^2\psi. \quad (5)$$

Hence, the governing equations for the vorticity and the temperature are

$$\frac{1}{\text{Pr}} \left( \frac{\partial \Omega}{\partial t} + \nabla \cdot (\Omega \mathbf{V}) \right) = \nabla^2 \Omega - \text{Ra} \frac{\partial \theta}{\partial x}, \quad (6)$$

$$\frac{\partial \theta}{\partial t} + \nabla \cdot (\theta \mathbf{V}) = \nabla^2 \theta + \varepsilon. \quad (7)$$

The boundary conditions (4a–c) are also imposed at the walls of the vessel.

When modelling a pipe as a line source, Beck *et al* (1988) used Green's functions and developed a transient solution to describe heat conduction around a single buried steam pipe inside a semi-infinite medium. The basic idea of modelling the wire as a source term in the energy equation is then to cast the source term  $\varepsilon$  as follows

$$\varepsilon = \delta(x_s - x) \delta(y_s - y), \quad (8)$$

where  $\delta(z)$  is the Dirac delta function. The integral of  $\delta(x_s - x) \delta(y_s - y)$  over  $x$  and  $y$  equals unit, if it includes  $(x_s, y_s)$ ; otherwise it is zero. Peyret (1990) used a similar technique but with an exponential decay of the intensity of the source term so that it could be represented accurately with spectral decomposition.

At steady state, the energy dissipated by the line heat source is lost at the boundaries. This requirement of equality leads to the non-dimensional steady state condition,

$$\int_{\mathbb{D}} \delta(x_s - x) \delta(y_s - y) dS = 1 = - \int_{\Gamma} \frac{\partial \theta}{\partial \mathbf{n}} d\Gamma, \quad (9)$$

where  $\mathbb{D}$  is the problem domain and  $\Gamma$  the boundary.



For adiabatic vertical walls, the heat flux through any horizontal plane  $y$ , defined as

$$\phi(y) = \int_0^1 \left( -\frac{\partial \theta}{\partial y} + V\theta \right) dx, \quad (10)$$

must satisfy the following conditions:

$$\phi(y^-) + \phi(y^+) = 1, \quad \forall y^- < y_s \quad \text{and} \quad \forall y^+ > y_s. \quad (11)$$

Finally, the depth of immersion  $H_s = A - y_s$  is defined as the vertical distance of the line source or the centre of the cylinder with respect to the top surface of the vessel.

### 3. Numerical procedure

#### 3.1 Numerical method for the finite-diameter cylinder

A finite element method was used for solving the governing PDEs in the primitive variable formulation, (1)–(3), with the boundary conditions given by (4a–d). While finite elements are CPU-time consuming, modelling a cylinder inside a rectangular vessel is straightforward. This is why a commercial fluid dynamics analysis package (FIDAP 1991) was employed.

A major difficulty in the application of the Galerkin finite element method for the incompressible Navier–Stokes equations is the elimination of spurious pressure modes. As a remedy for the checker-board pressure modes, a mixed-interpolation method was used, i.e. the interpolation of the pressure was a polynomial at least one order lower than that for the velocity. However, this method does not always provide accurate pressure fields, especially with bilinear interpolation for velocity and piecewise constant pressure (i.e., the  $Q_1/P_0$  finite element). The elimination of the pressure via a penalty function approach not only reduces the size of the linear system but eliminates the spurious pressures. The combination of the mixed interpolation method and the penalty formulation with exact (consistent) integration for the penalty term is one of the basic algorithms for the treatment of the velocity-pressure coupling of incompressible flows in FIDAP (1991). It is worth noting that the exact integration is equivalent to the reduced integration technique for the  $Q_1/P_0$  finite element and results in the same penalty matrix (Engelman *et al* 1982).

For the present computations, the four-node quadrilateral finite element was used for velocity and temperature variables with piecewise constant pressure and penalty function approximation. Integrals were evaluated exactly using one point Gaussian quadrature. At every selected Rayleigh number, the successive substitution method was employed for the first three iterations when solving the system of nonlinear equations, while the Newton–Raphson method was chosen after the third iteration. With these combinations, solutions converged smoothly to a 0.1% convergence criterion of the relative velocity and residuals within six to seven iterations for all steps. The selected penalty constant was  $10^{-8}$  for all cases and Rayleigh numbers.

Finite element meshes were built over a Cartesian  $33 \times 33$  grid and cylindrical mesh refinement was only used near the cylinder (figure 2); mesh refinement is needed in the vicinity of the cylinder when decreasing its size in order to maintain the accuracy of the results, the four-node quadrilateral element being straight-sided. The numbers of four-node quadrilateral elements and nodal points used for different diameters of

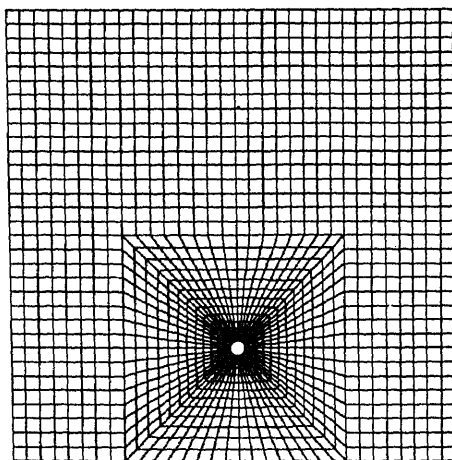


Figure 2. Finite element grid.

cylinder inside a square vessel are reported in table 1. Thus, 1152 four-node elements were used for the  $1/10$  diameter cylinder while 1664 elements were used for  $d = 1/100$ . Another consequence of mesh refinement was that, after the three successive substitution iterations, four Newton–Raphson iterations were needed to reach the convergence criterion for  $d = 1/100$  instead of 3 for greater diameters. Thus, the elapsed CPU time greatly increased for very small diameters.

A nine-node quadrilateral Lagrangian finite element for the velocity and temperature associated with linear discontinuous pressures, the  $Q_2/P_1$  finite element, was also tested (Deschamps & Saiac 1992). Some numerical tests of FIDAP have also been performed for some classical benchmarks by Sohn (1988) showing the efficiency of the  $Q_2/P_1$  finite element for the treatment of velocity–pressure coupling of incompressible fluid flow. This finite element has the major advantage of not presenting checker-board pressure modes. The penalty function method was preferred instead of the integrated method (for which the pressure variables contribute an additional degree of freedom and are built on linear continuous interpolation functions) because the memory storage required for fine meshes was found prohibitive and significantly

**Table 1.** Number of four-node elements according to cylinder diameter in a square vessel ( $H_s = 0.75$ ,  $Ra = 10^5$ ). Maxima are grid-values and the subscript 1 is for the point  $M_1(0.5, 3A/4)$ .

Diameter	Number of		Iterative		$\psi_{\max}$	$V_{\max}$	$U_{\max}$	$\theta_1$
	Nodal points	Elements	SS	NR				
1/10	1248	1152	3	3	5.34	42.04	22.86	0.133
1/20	1376	1280	3	3	5.65	45.30	23.80	0.137
1/50	1504	1408	3	3	5.83	46.98	24.36	0.139
1/70	1632	1536	3	3	5.87	47.88	24.99	0.140
1/100	1760	1664	3	4	5.90	48.25	24.58	0.140

SS: successive substitution method; NR: Newton–Raphson method

**Table 2.** Comparisons of some characteristic values for two different finite element discretizations and two Rayleigh numbers in a square vessel.

Ra	Element type	Number of		Iterative methods		$\psi_{\max}$	$V_{\max}$	$U_{\max}$	$\theta_1$
		Nodal points	Elements	SS	NR				
$10^5$	4-node	1632	1536	3	3	5.87	47.88	24.49	0.140
	9-node	1632	384	3	3	5.83	47.79	24.37	0.139
$10^7$	4-node	1632	1536		7	46.7	446.6	327.6	0.057
	9-node	1632	384		7	46.0	445.2	318.9	0.056

$d = 1/70$ ,  $H_s = 0.75$  (at  $Ra = 10^7$  a guess solution was used).

SS: successive substitution method; NR: Newton-Raphson method

Subscript 1 is for the point  $M_1$  (0.5, 3A/4)

increased the CPU time. Table 2 presents some local values in a square vessel for two different Rayleigh numbers. In order to avoid dramatic increases of the CPU time, about the same number of nodal points were used in both cases. Nevertheless, the CPU time was greater for the nine-node element, especially for the highest Rayleigh number (about 25% more). Consequently, all the runs were performed with the four-node element, the nine-node element showing unexpectedly slight accuracy improvement (Huyakorn *et al* 1978).

For transient problems, the time integration was performed by the combination of two accurate second-order techniques: the implicit trapezoid rule for the velocity and an explicit Adams-Bashforth formula for the pressure. If the solution at each time step is reasonably close to the solution at the previous time step, a one-step Newton-Raphson method can be used to solve the nonlinear system of algebraic equations. Typically, a value of  $\Delta t = 0.003$  was used at  $Ra = 10^7$  in a square vessel for a cylinder having a diameter of  $1/70$ . All the finite elements computations were performed on an superscalar workstation.

### 3.2 Numerical method for the line heat source

In the case of the local source term strategy, the vorticity and energy equations (5)–(7) were solved in transient form and the time integration was performed using an alternating directional implicit (ADI) splitting scheme. The vorticity equation was discretized by employing second-order central differences based on Taylor series expansions for all spatial derivatives, including the convective terms. The nodal points were located on a standard mesh. For the energy equation, a control-volume formulation with staggered grids and central differencing was retained in order to improve the overall energy balance. One layer of grid points outside each boundary was included to facilitate the application of the boundary conditions using quadratic extrapolations. For both vorticity and energy equations, the Thomas algorithm was employed to solve the tridiagonal systems of algebraic equations. On the other hand, finite difference equations for the stream function equation were solved by a direct method which uses a block-cyclic reduction process (Golub & Meurant 1983). From

the solution of the stream function equation, the wall vorticities were updated using an accurate second-order formulation based on Jensen's estimate in conjunction with Briley's formula (Roache 1982, p. 142). On account of the expected flow structure, uniform grids were used for all the computations discussed in the present paper. However, it should be noted that the main limitation of the direct method which was used is that the number of mesh points in one of the directions has to be chosen as a power of 2. Double precision computations were found necessary to compute accurately the threshold values and the characteristics of the oscillatory motions.

The computer code was validated with the steady and oscillatory solutions of a benchmark problem of natural convection of low-Prandtl fluids in differentially heated shallow cavities (Desrayaud *et al* 1990). Very good agreement was found both in the overall (general trends of unsteady flows) and pointwise comparisons (stream function and velocity maxima) with other computational methods: finite volume, finite element and spectral (Roux *et al* 1990).

Several common indicators of dynamics were employed for tracking convergence to an asymptotic state. Global indicators were calculated by using the discrete  $L_1$  or  $L_2$  norms on the flow data or on the flow change data. These are:

—the relative  $L_1$  norm of the stream function change per time step

$$\|\psi\|_1 = \Sigma |\psi^{n+1} - \psi^n| / \Sigma |\psi^{n+1}|;$$

—the total kinetic energy per surface unit:  $KE = (1/2D) \|\mathbf{V}\|_{d,2}^2$ ;

—the heat flux  $\phi$  through the top wall. In some cases, upper half heat fluxes, right  $\phi_r$  and left  $\phi_l$ , have been used.

—the  $L_2$  norm of the temperature per surface unit:  $PE = (1/D) \|\theta\|_{d,2}^2$ .

Since small-scale flow features can be masked by global indicators, a series of local variables, such as wall vorticity, velocity components or temperature were also recorded.

For the problem reported here, a number of tests were performed with various grid sizes and time steps to ensure accuracy and stability and to avoid spurious aperiodic flows (Desrayaud & Lauriat 1991). It was observed that the vorticity value in the vertical central plane was the most sensitive indicator, not only when symmetry-breaking transition occurred but also when the flow underwent a Hopf bifurcation.

Most of the computations were performed on an IBM 3090 600/VF vector computer. For typical cases, the vectorized performance (ratio of scalar to vectorial CPU time) was only slightly greater than two. The reason for this is mainly the difficulty of vectorizing the bloc-cyclic reduction solver. When using an ADI scheme to solve the Poisson equation of the stream function, the vectorized performance was over three. However, it is preferable not to introduce an internal iterative procedure when accurately transient motions are followed.

### 3.3. Convergence history

Calculations for the line heat source solution were done on three different regular grid structures in square vessels. Grid tests were conducted to ensure that the results were independent of both the grid density and the size of the source. Table 3 shows partial results from the tests for Rayleigh numbers below the critical value ( $Ra_c \cong 3.1 \cdot 10^7$ ). The local and overall data from these calculations differed by less than one

**Table 3.** Comparison of some characteristic values for different Rayleigh numbers in a square vessel and for a depth of immersion  $H_s = 0.75$  in the case of a line heat source. Subscript 1 is for the point  $M_1(0.5, 3A/4)$ .

Ra	Grid	$\phi$	$\theta_s$	$\psi_{\max}$	$V_{\max}$	$U_{\max}$	$\theta_1$
$10^5$	$33 \times 33$	0.688	0.281	6.13	49.70	24.80	0.140
	$65 \times 65$	0.687	0.376	6.14	50.52	25.01	0.141
	$129 \times 129$	0.687	0.480	6.14	50.73	25.09	0.141
$10^6$	$33 \times 33$	0.813	0.125	18.52	157.9	97.41	0.085
	$65 \times 65$	0.812	0.250	18.53	162.0	101.3	0.086
	$129 \times 129$	0.811	0.343	18.56	163.0	101.9	0.086
$10^7$	$33 \times 33$	0.886	0.102	46.05	418.9	290.3	0.052
	$65 \times 65$	0.881	0.157	46.45	438.9	316.3	0.055
	$129 \times 129$	0.879	0.231	46.60	444.5	322.3	0.055

percent for the two finest meshes. Note that maxima are grid-point values. For all the cases, the data were mostly grid-dependent around the heat source; since the heat is introduced into one control volume only, the temperature and flowfield in the immediate vicinity of the heat source cannot be grid-independent. The source temperature must tend to infinity as the area of the control volume tends towards zero. The decrease in the source temperature ( $\theta_s$ ) when the Rayleigh number is increased is due to the choice of the dimensionless variables. It is worth noticing that the results are satisfactory even for the coarsest mesh (within 5%), showing that stationary flow could be modelled using quite coarse meshes. Furthermore, two different regular grid structures were used to test the grid independence in a rectangular vessel of aspect ratio 2 (table 4). For these cases ( $H_s = 1$ ), the onset of periodic motion arises at low Rayleigh numbers and coarse grids thus give accurate results (within 1%) for stationary motions.

Extensive grid testing was also conducted for periodic motions. Test runs were made in square vessel at high Rayleigh numbers (table 5) and in a rectangular vessel at low Rayleigh numbers for various depths of immersion (table 6). It is seen that a strong dependency on the critical Rayleigh number is only found at high Rayleigh numbers, while a coarse grid gives accurate threshold and periodic motion at low Rayleigh numbers in the rectangular vessel whatever the depth of immersion. It demonstrated that, even if a weak dependence on mesh size is found for stationary

**Table 4.** Comparison of some characteristic values for symmetric and asymmetric flows in a rectangular vessel ( $A = 2$ ,  $H_s = 1$ ). Subscript 2 is for the point  $M_2(0.25, 3A/4)$ .

Ra	Grid	$\phi$	$\psi_{\max}$	$V_1$	$\theta_1$	$U_2$	$V_2$
$10^3$	$33 \times 49$	0.509	0.14	0.00	0.265	-0.152	0.258
	$65 \times 97$	0.509	0.14	0.00	0.265	-0.152	0.259
$5 \times 10^3$	$33 \times 49$	0.619	2.22	2.62	0.305	-2.21	2.91
	$65 \times 97$	0.619	2.21	2.62	0.306	-2.21	2.92
$10^4$	$33 \times 49$	0.637	1.46	0.00	0.342	-1.52	9.08
	$65 \times 97$	0.637	1.47	0.00	0.343	-1.53	9.13

**Table 5.** Onset of periodic motion in a square cavity for a line heat source ( $H_s = 0.75$ ).

Grid	Ra ( $\times 10^7$ )	Frequency
33 $\times$ 33	6.0	319.8 $\pm$ 0.8
65 $\times$ 65	3.2	308.2 $\pm$ 0.6
129 $\times$ 129	3.1	307.6 $\pm$ 0.6

flows at high Rayleigh numbers (table 3), critical values and frequencies exhibit strong grid dependencies as soon as the flow undergoes Hopf bifurcations (table 5).

Furthermore, to ensure the independence of the periodic motion with regard to the time step, the solutions were computed for two different time steps, but only for the two coarsest meshes (33  $\times$  33 and 65  $\times$  65) in the square vessel. Frequencies were found to agree within one percent. Because hundreds of thousands of iterations are needed to reach established periodic motion through supercritical Hopf bifurcations, the 129  $\times$  129 mesh flows were computed only once; computations for such a fine mesh are highly CPU-time consuming. Nevertheless, characteristics of the periodic motion are very similar to those found for the 65  $\times$  65 mesh, which gives us confidence in these results.

However, care should be taken in choosing the time step, since spurious secondary bifurcations (quasi-periodic solutions) may appear during the time integration of periodic flows. Figure 3 presents the effects of too large a time-step ( $\Delta t = 10^{-4}$  instead of  $5 \times 10^{-5}$ ) on the transient. Starting from a steady motion at  $Ra = 3 \times 10^7$  in a square vessel, the flow undergoes a periodic state of frequency  $f = 294.6$  at a reduced time  $t \approx 1.7$ . Therefore, using too large a time step yields at first an oscillatory solution with about the right frequency (table 5), but with a shortened transient motion, showing an amplification effect of the round-off errors. Beyond  $t \approx 8$ , this amplification effect produces a spurious aperiodic flow characterized by several low independent frequencies. This result shows the need to re-compute from time to time some oscillatory solutions using a sequence of time steps to check the validity of the oscillatory results.

**Table 6.** Frequencies for two uniform grids and for different depths of immersion in a rectangular vessel ( $A = 2$ ), line heat source model.

Depth of immersion ( $H_s$ )	Grid	Source position	Ra	Frequency ( $\pm 0.12$ )
1.50	33 $\times$ 49	17, 13	2.5 $10^4$	2.07
	65 $\times$ 97	33, 25	2.5 $10^4$	2.07
1.00	33 $\times$ 49	17, 25	3 $10^4$	5.61
	65 $\times$ 97	33, 49	3 $10^4$	5.62
0.75	33 $\times$ 49	17, 31	3 $10^5$	15.17
	65 $\times$ 97	33, 61	3 $10^5$	15.14

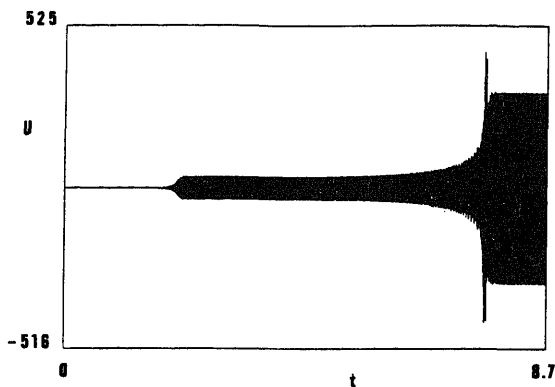


Figure 3. Time history from  $Ra = 3 \times 10^7$  to  $Ra = 3.1 \times 10^7$  of the horizontal component of the velocity at point  $(0.5, 2A/3)$  in a square vessel ( $65 \times 65$  mesh).

In transient motion leading to asymmetric steady flows, for depths of immersion greater than the width of the vessel (§4.2d), figure 4 shows the general trends for  $A = 2$  and  $H_s = 1.75$ . Starting from symmetric steady motion at  $Ra = 10^3$  to  $Ra = 6 \times 10^3$ , the global indicator on the flow change data decreases and becomes as small as  $10^{-13}$ . Then an opposite evolution occurs at  $t \geq 8$ , just after the break in the symmetry of the flow. At  $t \geq 25$ , the global indicator starts to decrease again. The time change is different for local indicators at the centre line of the vessel: the vorticity is equal to the zero machine for symmetric motion and increases continuously up to an asymptotic value as the time increases. Then, the flow reaches an asymmetric steady state motion (see figure 11 below).

Finally, figure 5 presents transient evolutions through a supercritical bifurcation like the ones arising when the depth of immersion is smaller than the width of the vessel (§§4.2b and 4.2c). Although the transition is easily detected in the case of subcritical Hopf bifurcations (since the oscillations arise with a finite amplitude), the transition from steady to oscillatory flows occurs through soft bifurcations such as the one displayed in figure 5. Such transitions can be the source of erroneous interpretations; the global indicators on the flow change data decrease to a level of the order of  $10^{-10}$ , while the local ones on the flow data stay at a relatively high level. In these cases, time integration must be pursued in order to show up a dramatic change. Thus, the usual indicators employed to stop the computations must only be used to scrutinize the transient behaviour of the flow in direct simulations of the route to chaos. Figures 4 and 5 show also that local indicators are more meaningful

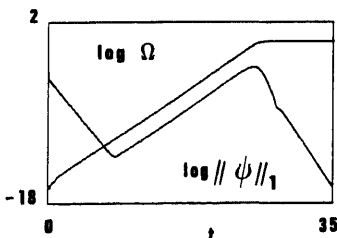


Figure 4. Time histories from  $Ra = 10^3$  to  $Ra = 6 \times 10^3$  of the logarithm of the vorticity at point  $(0.5, A)$  and of the  $L_1$  norm of stream function change in a rectangular vessel ( $A = 2$ ,  $H_s = 1.75$ ,  $33 \times 49$  mesh).

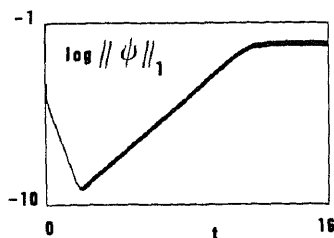


Figure 5. Time histories of the logarithm of relative  $L_1$  norm of stream function change for supercritical Hopf bifurcation,  $A = 2$ ,  $H_s = 0.75$ ,  $Ra = 3 \times 10^5$ .

than global ones, although local indicators must be chosen carefully according to the characteristics of the flow.

All these tests were performed for a line heat source owing to the fairly low computational cost. The numerical solutions of the Navier–Stokes equations written in the stream function–vorticity formulation and discretized by finite volume methods are much less time-consuming than finite element methods for solving the  $(u, v, p)$  equations, especially for transient motions.

## 4. Results

### 4.1 Comparisons between the two approaches

Figures 6 (a–c) show comparisons for steady state temperature and velocity profiles in a square vessel at  $Ra = 10^6$ . The depth of immersion is  $H_s = 0.75$  and the plots are for two cylinder sizes, of diameters  $1/20$  and  $1/70$ , and for the heat source term modelling with a  $65 \times 65$  mesh. The profiles for a  $1/100$ -diameter cylinder are not plotted here because no visible difference can be seen between them and those for the  $1/70$ -diameter. The temperature profiles on the vertical centreplane (figure 6a) show that the temperatures tend towards those of a line heat source when the diameter of the cylinder decreases, although some differences exist below the heat source. The vertical velocity profile in the vertical middle plane  $x = 0.5$  is shown in figure 6b. As can be seen, a non-zero velocity exists at the line source when it is modelled as a heat source term. However, the agreement between the velocity profiles for the two approaches is rather good in the ranges  $0 \leq y \leq 0.2$  and  $0.4 \leq y \leq 1$  in the case of a  $1/70$ -diameter cylinder. Finally, the influence of the cylinder diameter on the vertical velocity profile in the horizontal plane  $y = A - H_s$  (i.e. the plane passing through the centre of the cylinder) is displayed in figure 6c. While two velocity maxima are seen on both sides of the cylinder, owing to the no-slip boundary conditions, and only one for the line heat source, the differences between the two models are small on the major part of the profiles. Obviously, the discrepancies in the vicinity of the cylinder decreases when decreasing the cylinder diameter.

Figure 7a shows the relative differences for the maximum of the stream function,  $\psi_{\max}$ , and for the heat flux through the top wall,  $\phi$ , as a function of the Rayleigh number when the cylinder diameter is increased from  $d = 1/100$  to  $1/10$ . The solid lines are for  $\psi_{\max}$  while the dashed lines are for  $\phi$ . These are the 1, 2 and 5% iso-difference curves, the reference value being the solution for a cylinder of  $1/100$ -diameter. For  $Ra \leq 10^5$ , the motion is in the so-called conduction or transition regimes. More than 65% of the heat being transferred by conduction to the bottom



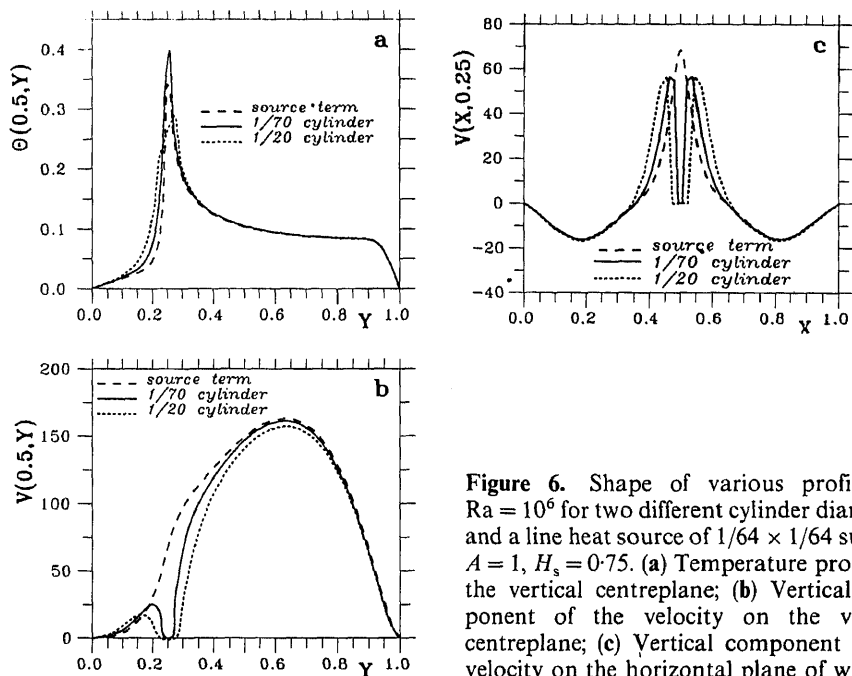


Figure 6. Shape of various profiles at  $Ra = 10^6$  for two different cylinder diameters and a line heat source of  $1/64 \times 1/64$  surface.  $A = 1$ ,  $H_s = 0.75$ . (a) Temperature profile on the vertical centreplane; (b) Vertical component of the velocity on the vertical centreplane; (c) Vertical component of the velocity on the horizontal plane of wire.

wall at  $Ra = 10^4$ , the temperature field is not very sensitive to the size of the cylinder. For  $d = 1/60$  and  $1/20$ , the relative differences in  $\phi$  are then close to 1% and 5% respectively. On the contrary, larger discrepancies appear for the stream function because of the low velocity field. The cylinder is then an obstacle to the fluid motion, since the velocities are small: at  $Ra = 10^4$ , the differences in the stream function for  $d = 1/60$  is about 4% and reaches 17% for  $d = 1/20$ . Thus, the flow is greatly affected

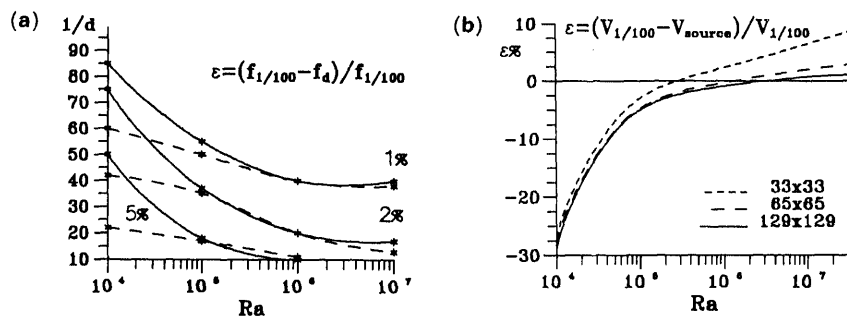


Figure 7.  $A = 1$ ,  $H_s = 0.75$ . (a) Curves of relative iso-difference for the cylinder model. Solid lines: stream function maximum, dashed lines: upper heat flux. (b) Curves of relative difference of  $V_{max}$  for different mesh sizes of the source term model. The reference is the  $d = 1/100$  cylinder.

by the size of the cylinder in the conduction and transition regimes. On the other hand, the discrepancies are smaller in the boundary layer regime because heat is then mainly convected downstream by the plume. It appears that an asymptotic value of the diameter is reached for a given relative difference:  $1/40$  for 1% and  $1/20$  for 2%. It occurs for a Rayleigh number close to  $10^6$ , i.e. at the beginning of the boundary layer regime.

Figure 7b presents the evolution of the relative difference as a function of the Rayleigh number for the source term model and for the three meshes used:  $33 \times 33$ ,  $65 \times 65$  and  $129 \times 129$ . The comparisons are for the maximum of the vertical component of the velocity, which is always on the vertical centreplane of the vessel. According to  $Ra$ , the location of the maximum velocity moves between  $y = 0.5$  and  $0.7$ . Here again, the reference value is the maximum vertical velocity in the vertical plane  $y = 0.5$  for a cylinder of diameter  $d = 1/10$  and the same conclusions as for figure 7a can be drawn. The largest discrepancies are about 30% for  $Ra = 10^4$ , whatever the mesh size. This relative difference decreases when the Rayleigh number increases and remains small for the two finest meshes, even at  $Ra = 3 \cdot 10^7$ . For the coarsest mesh ( $33 \times 33$ ), it reaches 10%. Nevertheless, it is worth noting that the curves for the three meshes exhibit the same trends.

The relative differences between the maxima of the vertical velocity component in the vertical middle plane and in the horizontal plane passing through the cylinder axis are reported in table 7 for various cylinder diameters and depths of immersion. On account of the symmetry of the flow, only one velocity maximum needs to be considered in the horizontal plane while two velocity maxima have to be considered in the vertical middle plane: one below the heated cylinder and one above, into the thermal plume. For all the data reported in table 7, the reference values are the maximum velocities at the same planes for a cylinder with a diameter  $d = 1/100$ . The relative differences between the velocity maxima in the horizontal plane are denoted  $H$  and are reported in the left-hand columns of table 7 for the three depths of immersion investigated. The relative differences between the vertical velocity maxima are denoted  $VA$  and  $VB$ , above and below the cylinder respectively. All the computations were carried out for  $Ra = 10^6$ . At this Rayleigh number, the plume flow is at the beginning of the boundary layer regime. For higher Rayleigh number, the differences are of the same order while they are greater for lower Rayleigh number because the motion is weaker.

As can be seen, the relative differences in the vertical centreplane (see  $VA$ - and  $VB$ -columns) depend mainly on the diameter of the cylinder and are almost independent of the depth of immersion, except for  $d = 1/10$ . These differences are a little less below the cylinder than above. However, the size of the cylinders greatly affects the motion, especially for  $d \geq 1/30$ . On the contrary, the relative differences in the horizontal plane of the cylinder are not only affected by the size of the cylinder but also by its location. Table 7 shows that these differences increase strongly with the depth of immersion.

The distances from the centre of the cylinder to the location of the three velocity maxima are also reported in table 7,  $\Delta x$  being on the horizontal plane and  $\Delta y$  on the vertical centreplane. At constant depth of immersion, these values are almost constant provided  $d < 1/20$ . For greater diameters, the maxima are slightly shifted away. This shows that the influence of the size of the cylinder on the flowfield is limited to the very near surroundings of the cylinder.

For periodic flows, computations were carried out for only one cylinder diameter,

**Table 7.** Relative differences in percent between maximum of the vertical component of the velocity for various depths of immersion.  $Ra = 10^6$   $A = 1$ . $\Delta x, \Delta y$  are the positions of the maxima from the centre of the cylinder

$d$	$H_s = 0.25$			$H_s = 0.5$			$H_s = 0.75$		
	$H(\%)$	$VB(\%)$	$VA(\%)$	$H(\%)$	$VB(\%)$	$VA(\%)$	$H(\%)$	$VB(\%)$	$VA(\%)$
1/10	22.9	28.6	57.4	37.9	14.2	53.8	62.3	13.5	46.4
$\Delta y$		0.064	0.340		0.075	0.325		0.062	0.312
$\Delta x$	0.090			0.103			0.092		
1/20	9.4	17.6	30.5	22.2	12.9	29.9	37.8	14.5	29.1
$\Delta y$		0.41	0.314		0.049	0.299		0.035	0.286
$\Delta x$	0.063			0.075			0.061		
1/30	5.4	11.7	19.1	12.6	9.3	18.6	25.6	11.8	21.4
$\Delta y$		0.041	0.314		0.026	0.299		0.025	0.275
$\Delta x$	0.063			0.075			0.061		
1/50	2.2	5.7	8.8	5.2	4.8	8.4	10.6	6.8	10.5
$\Delta y$		0.041	0.314		0.026	0.299		0.025	0.275
$\Delta x$	0.063			0.075			0.048		
1/70	0.9	2.6	3.9	2.2	2.3	3.7	5.6	4.0	5.7
$\Delta y$		0.041	0.314		0.026	0.299		0.025	0.275
$\Delta x$	0.063			0.075			0.048		

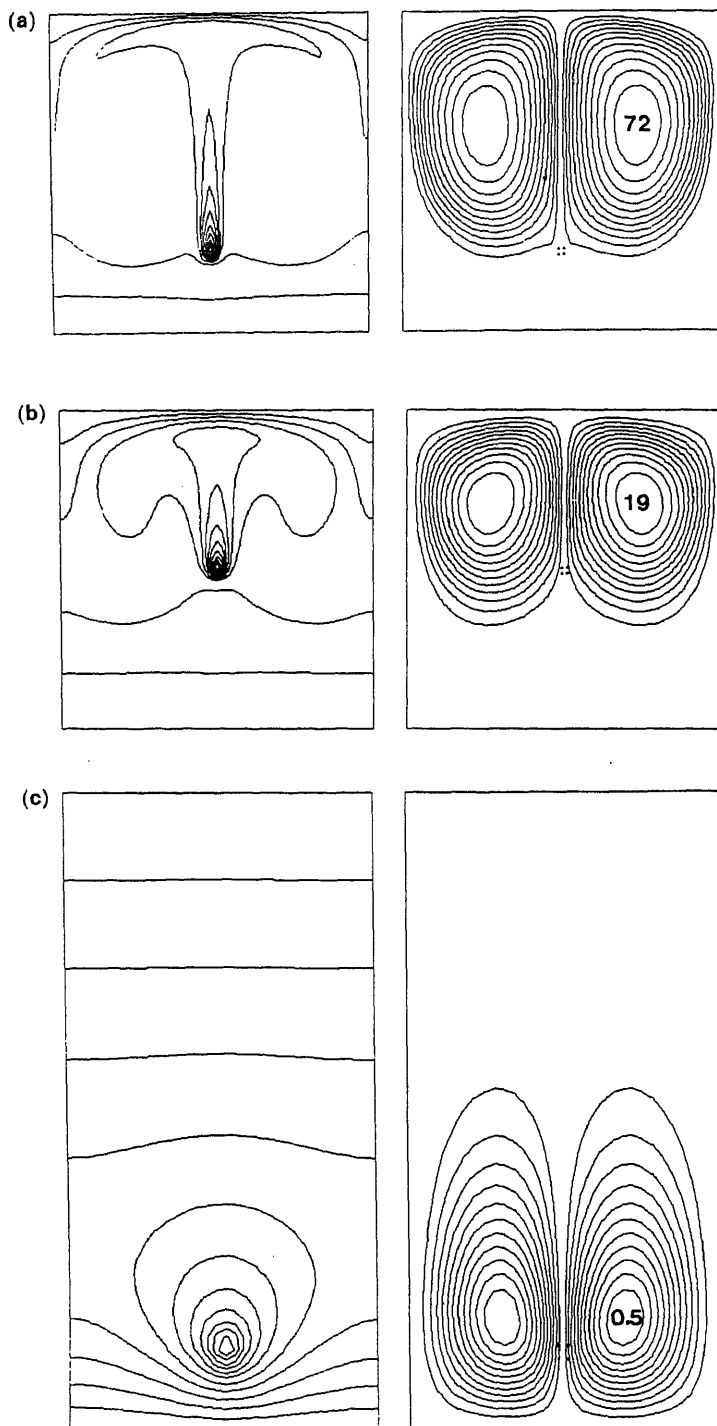
$d = 1/70$ , because transient simulations at high Rayleigh number are highly CPU-time consuming when using a commercial finite-element package. According to the finite element grid used, the Rayleigh number at which instabilities occur was found to be very close to  $Ra = 7 \times 10^7$  and the system is then attracted to a limit cycle representing a periodic motion of frequency  $f = 357 \pm 38$ . This Rayleigh number and frequency are close to those obtained with the source term formulation.

#### 4.2 Route to chaos of a buoyant plume in rectangular vessels

On account of the rather good agreement between the solutions computed at high Rayleigh numbers by using one or the other of the two approaches, all the results presented in the following were obtained through the source term formulation.

**4.2a Base flow simulations:** Typical results for the streamlines and isotherms are shown in figures 8a–c for various depths of immersion of line sources in the central vertical plane of the vessels. Maxima of the stream function are also given. All the simulations have been performed for air-filled vessels ( $Pr = 0.71$ ).

Table 8 gives the lower and upper bounds on the critical Rayleigh numbers. These bounds are for the highest value at which a steady-state motion was found to exist ( $Ra_1$ ) and the smallest value at which the flow was found unsteady ( $Ra_2$ ). The frequencies reported in table 8 are for the  $Ra_2$ -values. The different types of bifurcation occurring in the configurations considered are also given. Steady-state isotherms and streamlines in square and rectangular vessels for Rayleigh numbers just below the first bifurcation point are shown in figure 8. It can be seen that a recirculation flow is induced in which hot fluid rises with the plume above the line source, is cooled downstream and then descends along both sides of the vessel. The flow fields are



**Figure 8.** Steady-state isotherms (left) and streamlines (right) for various flows and Rayleigh numbers just below the critical values,  $Pr = 0.71$ . (a)  $Ra = 3 \times 10^7$ ,  $A = 1$ ,  $H_s = 0.75$ ; (b)  $Ra = 1.9 \times 10^6$ ,  $A = 1$ ,  $H_s = 0.50$ ; (c)  $Ra = 5 \times 10^3$ ,  $A = 2$ ,  $H_s = 1.75$ . The values given refer to stream function maxima.

**Table 8.** Critical Rayleigh numbers, frequencies and types of Hopf bifurcation at various depths of immersion for  $A = 1$  (a) and 2 (b) and for various aspect ratios at  $H_s = 0.75$  (c).(a)  $A = 1$ 

$H_s$	Bounds $Ra_1$ & $Ra_2$	$f$	Type of Hopf bifurcation
0.75	$3.0\text{--}3.1 \times 10^7$	294.2	Supercritical
0.50	$1.5\text{--}1.6 \times 10^6$	30.4	Supercritical
0.25	$3.1\text{--}3.2 \times 10^6$	38.5	Supercritical

(b)  $A = 2$ 

$H_s$	Bounds $Ra_1$ & $Ra_2$	$f$	Type of Hopf bifurcation
1.75	$3.00\text{--}3.02 \times 10^4$	0.2	Subcritical
1.50	$2.1\text{--}2.2 \times 10^4$	1.7	Subcritical
1.25	$1.0\text{--}1.5 \times 10^4$	1.9	Subcritical
1.00	$2.0\text{--}2.5 \times 10^4$	5.3	Supercritical
0.75	$2.5\text{--}3.0 \times 10^5$	15.0	Supercritical

(c)  $H_s = 0.75$ 

$A$	Bounds $Ra_1$ & $Ra_2$	$f$	Type of Hopf bifurcation
2.00	$2.5\text{--}3.0 \times 10^5$	15.0	Supercritical
1.75	$3.0\text{--}4.0 \times 10^5$	17.2	Supercritical
1.50	$8.0\text{--}9.0 \times 10^5$	25.6	Supercritical
1.25	$1.0\text{--}1.5 \times 10^6$	33.2	Supercritical
1.00	$3.0\text{--}3.1 \times 10^7$	294.2	Supercritical

characterized by mirror symmetry about the vertical centreline. It is seen from figure 8a ( $A = 1$ ,  $H_s = 0.75$ ) that for line sources near the bottom wall, a strong steady fluid circulation occurs above the line source while a relatively stagnant layer of fluid is seen below. Obviously, the temperature gradients inside the bottom region strongly depend on the thermal boundary conditions applied at the bottom wall while the flow and temperature fields in the upper region of the vessel are weakly affected by this thermal boundary condition provided the Rayleigh number is large enough (Lauriat & Desrayaud 1990). Similar features are seen in figure 8b for smaller depth of immersion,  $H_s = 0.5$ , but the bottom stagnant layer of fluid which is stably stratified now extends over one-third of the vessel approximately. On the other hand, for large depths of immersion in rectangular vessels as shown in figure 8c, the plume does not reach the top wall and there is an unstably stratified layer of stagnant fluid above the plume. It should be noted that the fluid circulation is very weak owing to the low Rayleigh number. The flow is in a conductive state, as confirmed by the quasi-circular isotherms around the source.

These three flow patterns give way to three different routes to chaos (Desrayaud & Lauriat 1993). The first scenario, studied in the next sub-section, can be found only if the layer of fluid below the line source is small enough and if the plume reaches the top of the vessel. This happens only in vessels of small aspect ratio ( $A \leq 1$ ). The resulting periodic motion can be viewed as the natural swaying motion of confined plumes in the sense that the instabilities are neither driven by a stable layer of fluid

at rest below the source as in figure 8b nor triggered by an unstable layer of fluid above the plume as in figure 8c. The two other scenarios are called penetrative convection, since the convecting plume is bounded below by a conducting layer of fluid, and Rayleigh–Bénard-like convection, since the underlying mechanism of the onset of instabilities is the destabilization of a motionless upper layer. These two mechanisms are respectively studied in §§ 4.2c and 4.2d below.

Thus, contrary to what has been found in many studies on freely-rising thermal plume, interactions of the plume with the top and side walls play a dominant role in the flow structure.

**4.2b Natural swaying motion in a square vessel:** In this sub-section the numerical results discussed are for vessels of square cross-section, the line source being near the cold bottom wall ( $H_s = 0.75$ ) as in figure 8a.

**Periodic motion** – For Rayleigh numbers lower than  $Ra_c = 3 \times 10^7$ , the system is attracted to a fixed point, representing steady motion. The bifurcation leads to a system which is then attracted to a limit cycle indicating periodic motion.

The onset of unsteady solutions is due to the presence of a supercritical Hopf bifurcation point. Indeed, for such a bifurcation, the amplitude of the perturbation for slightly supercritical Rayleigh numbers evolves like  $(Ra - Ra_c)^{0.5}$ . This feature has been used to accurately determine the value of the critical point, which has been done from linear extrapolation of zero oscillation amplitude occurring close to the presumed threshold value. This gives a critical Rayleigh number very close to  $3 \times 10^7$ . A further feature of a Hopf bifurcation is the existence of a relationship between the dimensionless period of the oscillations and the Rayleigh number in the vicinity of the bifurcation point. By calculating the angular Brunt–Väisälä frequency associated with the plume for high Rayleigh numbers, it has been demonstrated that the frequency must be proportional to  $(Ra \text{ Pr})^{0.5}$  and should be nearly constant near the threshold (Desrayaud & Lauriat 1993). This is well supported by the results of table 9 which shows the route to chaos, i.e. the nature of the bifurcation points and the associated frequencies.

A second frequency  $f_2$  (given in parenthesis in table 9) appears at  $Ra = 3.4 \times 10^7$  during the transient evolution but vanishes for a time unit greater than one, meaning that the frequency  $f_2$  has at least eight orders of magnitude less power; moreover, these two frequencies are incommensurate.

During one period, the symmetry with respect to the vertical centre plane is respected. This finding was experimentally recorded by Yosinobu *et al* (1979) in the case of a buoyant plume in air. The general pattern of the mean temperature and stream function fields are the same as those presented in figure 8a, but with higher isovalues. It can be noted that the upper part of the plume has the same symmetrical motion, once to the left, once to the right of the cavity. Hot and cold fluctuations grow simultaneously on each side of the source and a circulation of alternately hot and cold fluctuations arises in the two halves of the vessel. The instabilities are first confined within the plume where they are amplified and within the horizontal boundary layer near the top wall, then they move downward and back to the heat source. Therefore, instead of a swaying motion with sinusoidal wavelength as for a freely rising plume, we observed two counter-rotating circulations of fluid with alternating hot and cold spots. The motion of the plume itself is rather weak and detached blobs arise in the upper horizontal extents of the plume and sink along the

**Table 9.** Route to chaotic motion in a square vessel with a line source at  $H_s = 0.75$ 

$10^{-7} Ra$	Description	Frequency	$(Ra/f)^{1/2}$
3.00	Supercritical Hopf bifurcation		
3.10	P1	$f_1 = 294.2$	18.93
3.20	P1	$f_1 = 299.7$	18.87
3.30	P1	$f_1 = 305.2$	18.82
3.40	Transient QP2, then P1	$\left\{ \begin{array}{l} f_1 = 310.1 \\ (f_2 = 94.4) \end{array} \right.$	18.80
3.50	Transient QP2, then P1	$\left\{ \begin{array}{l} f_1 = 314.8 \\ (f_2 = 96.1) \end{array} \right.$	18.79
3.60	Transient QP2, then P1	$\left\{ \begin{array}{l} f_1 = 319.4 \\ (f_2 = 97.7) \end{array} \right.$	
3.70	P2T (weak $f_0$ )	$\left\{ \begin{array}{l} f_1 = 323.5 \\ (f_2 = 99.2) \end{array} \right.$	
3.80	P2T (weak $f_0$ )	$\left\{ \begin{array}{l} f_1 = 328.2 \\ f_2 = 101.1 \end{array} \right.$	
3.85	Transient P2T, then I		
3.90	Transient P2T, then I		
4.00	I		

P1 periodic state; QP2 quasi-periodic state with 2 incommensurate frequencies; P2T periodic state on a 2-torus; I intermittent state.

vertical adiabatic surface. Urakawa *et al* (1983) experimentally found identical behaviour in spindle oil but with a much stronger motion of the plume, especially just above the line source.

**Two-frequency locked state** – For  $3.7 \times 10^7 \leq Ra \leq 3.8 \times 10^7$ , the motion smoothly becomes a periodic, two-frequency locked state involving the  $f_1$  and  $f_2$  frequencies. The asymptotic state is then a limit cycle on a 2-torus of small cross-section. The phase portrait reveals that the trajectories are confined to a finite number of threads (figure 9a). The Poincaré section confirms this behaviour since 13 distinct group points are alternately visited in turn: for 13 rotations about its larger dimension, the trajectories pass four times around the smaller dimension (figure 9b). Thus, the rotation number (or the frequency-locking ratio) is  $r = 4/13$  and the fundamental frequency equals  $f_0 = f_1$ ,  $13 = f_2/4$ . This is well supported by the frequency values reported in table 9. Indeed,  $f_1/f_2 = 3.26 \pm 0.03 \cong 13/4$ . Moreover, the  $f_2$ -frequency has only a weak influence on the whole motion, confined in the centre of the cells. It should also be noticed that the nonlinearities are weak since the contribution of the low-order mixing components ( $f_1, f_2$ ) is small. Simulations have been carried out up to 150,000 time-steps ( $t > 3$ ), and no established quasi-periodic motion ( $f_1, f_2$  incommensurate) has ever been found.

**Chaos** – An intermittently chaotic state arises from the previous frequency-locked state. At irregular times and for irregular durations, the periodic laminar motion is interrupted by non-periodic 'bursts'. However, the characteristic of the frequency-locked state with locking ratio 4/13 is maintained in the laminar windows. As a result,

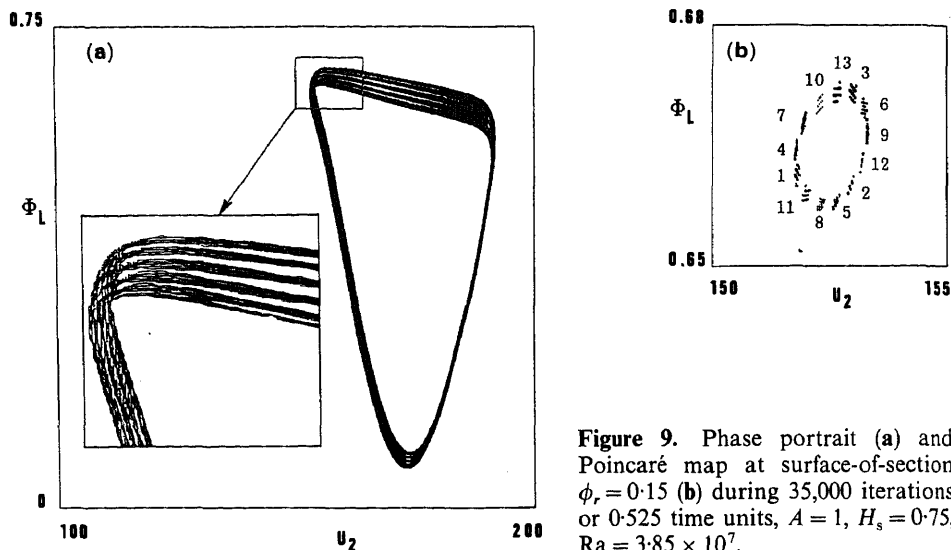


Figure 9. Phase portrait (a) and Poincaré map at surface-of-section  $\phi_r = 0.15$  (b) during 35,000 iterations or 0.525 time units,  $A = 1$ ,  $H_s = 0.75$ ,  $Ra = 3.85 \times 10^7$ .

the spectrum exhibits broadband noise although relatively sharp spectral peaks still exist for all frequency multiples of the 4/13 locked state (Lauriat & Desrayaud 1990).

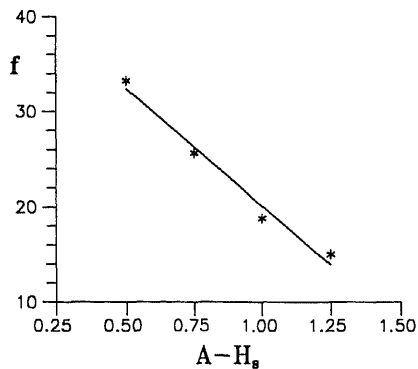
From their experiments, Forstrom & Sparrow (1967) reported turbulent bursts at the beginning of the transition between the laminar and turbulent states (Note that Yosinobu *et al* (1979) did not observe such phenomena). In the transition regime, Bill & Gebhart (1974) and Noto *et al* (1982) recorded transits from a turbulent state back to a laminar one, which seems like an intermittency phenomenon.

It could be concluded that a chaotic motion arises through a type-I intermittent transition. This type of intermittency is characterized by bursts of equal magnitude, periodic windows of identical frequencies, and near the transition, the lengths of these windows vary in proportion to  $(Ra - Ra_c)^{-0.5}$ . The sequence of instabilities leading to non-periodic flows is also described on table 9. The periodic two-frequency locked state has been abbreviated at P2T (periodic motion on a 2-torus).

**4.2c Penetrative convection:** As can be seen in figure 8b, there is a stable layer of fluid at rest at the bottom of the vessel for small depths of immersion although the convective motion extends slightly below the line source. Computations have been carried out in square and in rectangular vessels. For a square vessel, penetrative convection occurs if the depth of immersion is such that  $H_s \leq 0.5$  while for rectangular vessels it happens when the depth of immersion is smaller than the width of the cavity (i.e.,  $H_s \leq 1$ ) (Desrayaud & Lauriat 1993).

The bifurcation points are supercritical Hopf points with low frequency (cf table 8a,  $H_s \leq 0.5$  and table 8c,  $A \geq 1.25$ ) since the amplitude increases roughly as the square root of the distance to the bifurcation point and the modified period is almost constant. The low values of the frequencies found can be explained by the fact that the plume has to set in motion the fluid below the source. This is illustrated in





**Figure 10.** Variation of the frequency at critical points versus the height of the layer below the source for various aspect ratios and a fixed depth of immersion  $H_s = 0.75$ .

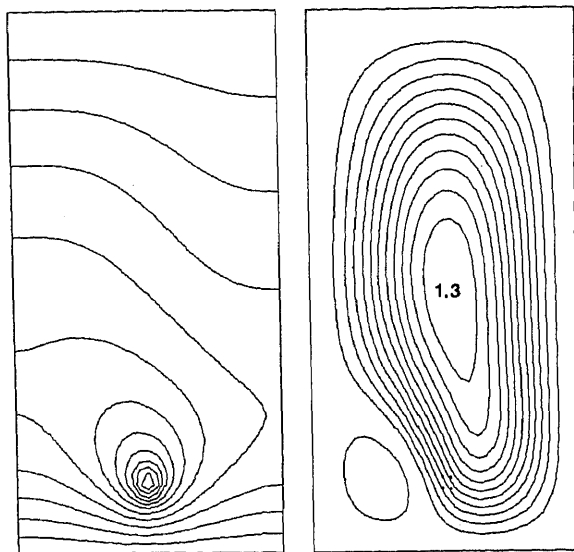
figure 10 which shows a linear variation of the frequency of the critical points (given in table 8c) with the height of the layer below the source ( $A - H_s$ ) for a constant depth of immersion,  $H_s = 0.75$ . Such a result clearly demonstrates that penetrative convection is the main phenomenon driving these instabilities. Some other cases in rectangular vessels and for depths of immersion  $H_s \leq 1$  were investigated and the same trends were found.

In a square vessel and for a depth of immersion  $H_s = 0.75$ , a periodic motion arises at  $Ra_c = 6 \times 10^6$  and is characterized by a large contribution from the two first harmonics. The flow then undergoes a second bifurcation into a limit cycle on a 2-torus. A weak frequency  $f_2$ , ten times smaller than the fundamental one  $f_1$ , appears at  $Ra = 1.31 \times 10^7$  and a two-frequency locked state motion with a rotation number  $r = f_2/f_1 = 1/10$  is obtained. The spectra of the trajectories exhibit a large number of sidebands around the fundamental peak and its harmonics resulting from a strong nonlinear interaction between the two-locked frequencies,  $f_1$  and  $f_2$ . As the Rayleigh number is increased, these sidebands develop further, both in number and amplitude and compete with one another leading to the thickening of the torus. Just after the onset of chaotic behaviour the sidebands grow throughout the spectrum, the underlying envelop being broadband, despite the sharpness of the fundamental frequency  $f_1$  and its two first harmonics. Thus, this results in a fast transition to a fully chaotic spectrum.

**4.2d Rayleigh-Bénard-like convection:** For rectangular vessels and depths of immersion greater than the width of the vessel ( $H_s > 1$ ), very different bifurcations occur.

**Pitchfork bifurcation** – On increasing the Rayleigh number up to  $Ra = 6000$ , the symmetric two-cell pattern shown in figure 8c evolves towards an asymmetric one-cell pattern as in figure 11. The plume is deflected towards one vertical adiabatic wall, either left or right depending on the round-off errors generated during the computations. These two steady-state mirror image solutions characterize a pitchfork bifurcation (Desrayaud & Lauriat 1993).

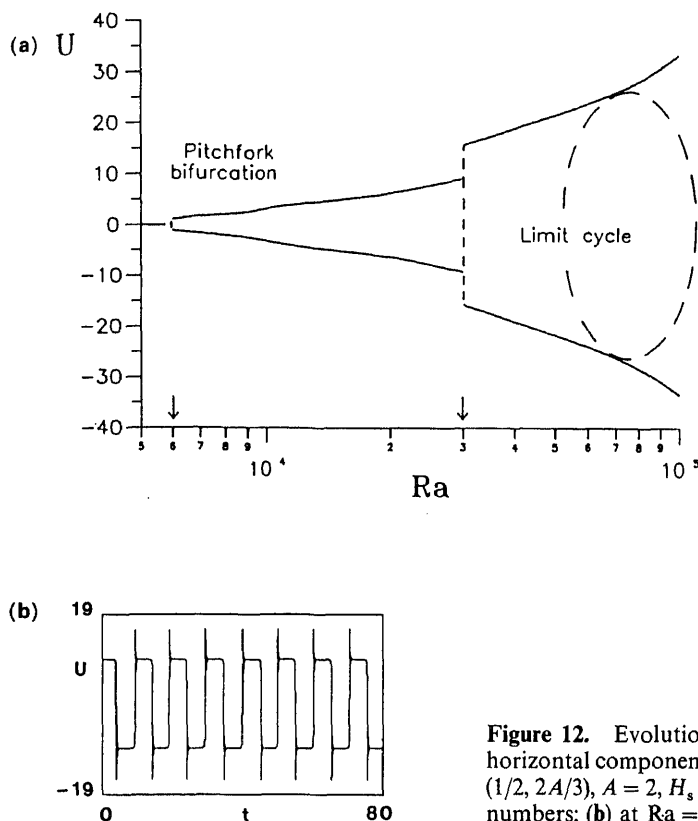
Contrary to what has been found for free laminar plumes, the destabilization of the two-cell flow comes from the unstably stratified layer of fluid at rest above the plume which appears when the plume does not reach the top of the vessel. Thus, Rayleigh-Bénard-like convection can arise in this upper layer when the Rayleigh



**Figure 11.** Isotherms and streamlines for stationary asymmetric flow,  $A = 2$ ,  $H_s = 1.75$ ,  $Ra = 6 \times 10^3$ . The values given refer to stream function maxima.

number is high enough, giving way to one-cell flow which spreads out in the vessel. The potential for multiplicities of steady-state mirror solutions is the result of nonlinearities of the governing equations. Similar behaviour has been observed by Hasnaoui *et al* (1990) for natural convection above an array of open cavities heated from below when the height of the vertical adiabatic confining walls is high enough.

*Subcritical Hopf bifurcation* – As the Rayleigh number is increased further, a sustained oscillatory convection is obtained through a subcritical Hopf bifurcation. Convenient variables to describe the temporal evolution of the flow are local variables on the centreplane of the vessel. Figure 12a presents the evolution of the horizontal velocity component  $U$  at point  $M_1(1/2, 2A/3)$  for various Rayleigh numbers and for a depth of immersion  $H_s = 1.75$ . Below the pitchfork bifurcation point, which occurs at  $Ra = 6 \times 10^3$ , its value is zero due to the flow symmetry. Above it, the velocity component can take two opposite values. For higher Rayleigh numbers, the motion undergoes a second bifurcation at  $Ra \cong 3.02 \times 10^4$  into a limit cycle and the value oscillates periodically between the two mirror values of the Pitchfork bifurcation. Thus, oscillation onset is with finite amplitude which defines a subcritical Hopf bifurcation. On decreasing the Rayleigh number from  $Ra = 3.02 \times 10^4$  to  $3.01 \times 10^4$  the flow became steady again. Consequently, no hysteresis effect was found. As can be seen in figure 12a, the value of the horizontal component of the velocity jumps at the Hopf bifurcation point. The underlying phenomenon is clearly depicted in figure 12b which presents the time history of this variable for a Rayleigh number just above the onset of oscillations. Plateaus of opposite values which are those of the pitchfork bifurcation are periodically reached. The plume sways abruptly, briefly overshoots the other mirror flow solution due to its own inertia and becomes stable over a long period of time. At  $Ra = 3.02 \times 10^4$  the period is equal to about 5 ( $f \cong 0.2$ ). This is why the frequencies are so small for  $H_s > 1$  (table 8b). On increasing



**Figure 12.** Evolution of the amplitude of the horizontal component of the velocity  $U$  at point  $(1/2, 2A/3)$ ,  $A = 2$ ,  $H_s = 1.75$ : (a) versus Rayleigh numbers; (b) at  $Ra = 3.02 \times 10^4$ .

the Rayleigh number, the solution continues to oscillate between two mirror solutions and the frequency is increased, the plateaus being shorter.

At higher  $Ra$ , it seems that the temporal behaviour of the oscillations becomes chaotic. The leading Lyapounov exponent (LLE) estimate, which measures how unstable a given flow history is, has been calculated (Wolf *et al* 1985). Its sign provides a qualitative picture of the system's dynamics. The positive sign of the LLE implies the existence of chaotic behaviour at least for  $Ra = 8 \times 10^5$ . However, the regime is difficult to map accurately owing to limitations in the numerical resolution. Indeed, due to the very low value of the frequency, long time integrations are needed to correctly observe transitions and to obtain accurate power spectra. Nevertheless, it does not seem that any period doubling scenario exists after the pitchfork bifurcation as for the classical Feigenbaum scenario.

**4.2e Comparisons with experimental results:** Igarashi (1978) experimentally found that the swaying frequency is proportional to the 0.4 power of Rayleigh number for heated wires placed concentrically along the axes of horizontal rectangular chambers and thermal plumes reaching the ceiling of the vessel in a laminar state. Frequency correlations for various aspect ratios are depicted in figure 13. Three distinct groups of frequencies have been found as predicted experimentally by Igarashi (1978), one for  $A \geq 1.8$ , another for  $0.7 \leq A \leq 1.5$  and the last for  $A \leq 0.6$ . Owing to the very high

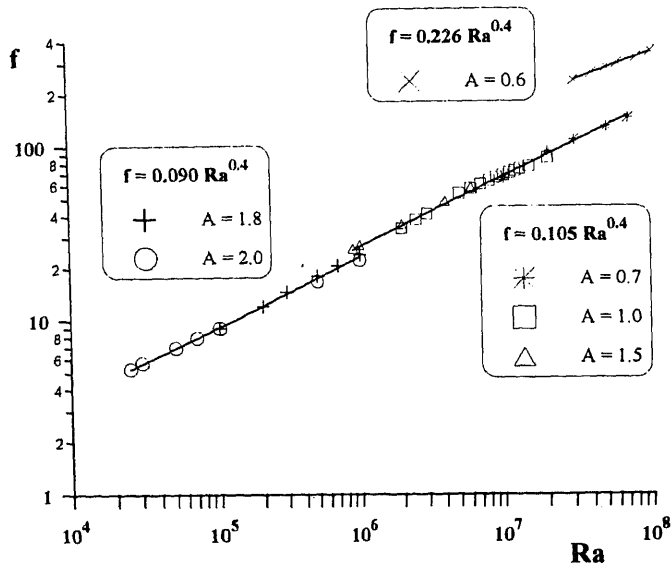


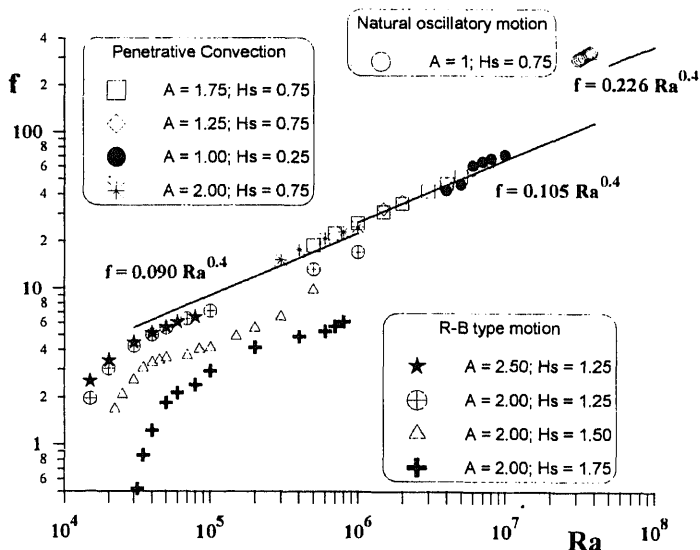
Figure 13. Evolution of the frequency for various aspect ratios and line heat source located at cavity mid-point.

Rayleigh number at which periodic motions appear for  $A \leq 0.6$ , computations have been made only for one aspect ratio, i.e.  $A = 0.6$ . For each numerical frequency, the value of  $f/Ra^{0.4}$  has been calculated and the constants of the three correlations shown in figure 13 are the arithmetic mean values of each group of data. As can be seen, all the data are well correlated. The maximum discrepancies are lower than 5% except at  $A = 0.7$  where the differences reach 10%.

The 0.4 power value has been recently confirmed by Noto & Matsumoto (1986) and Noto (1989). The aspect ratio of their experimental set-up was  $A = 1.25$  with a dimensionless depth of immersion  $H_s = 0.75$ . For the same geometrical parameters, a numerical frequency correlation has been computed,  $f = 0.109 Ra^{0.4}$  with discrepancies lower than 3%.

All these results give a good degree of confidence in the present results. It must be noted that the thermal boundary conditions play a minor role only. Indeed, in the experimental apparatus of Igarashi (1978) and Noto (1989), the temperatures of the ceiling, the bottom and the four sidewalls were isothermally controlled while our numerical boundary conditions are imposed by the top and bottom temperatures and vertical adiabatic walls.

Do the three correlations experimentally found by Igarashi (1978) correspond to the three oscillatory regimes numerically determined in the present study? To answer this question, all the numerical frequencies obtained in the present study are shown in figure 14, except those used to build the three correlations, i.e. for  $H_s = A/2$  (figure 13). This has been done to give a legible picture, and the correlations are reported in figure 14 as full lines. It is readily seen that the correlation  $f = 0.105 Ra^{0.4}$  is for the penetrative convection regime; all the numerical frequencies corresponding to this regime agreed very well with this correlation. The other correlation,  $f =$



**Figure 14.** Comparison of the frequencies for various aspect ratios and depths of immersion given in figure 13.

$0.090 Ra^{0.4}$ , which is close to the previous one, also corresponds to the penetrative convection regime. In fact, it is a particular case of this regime for which there is competition between two modes, the Rayleigh–Bénard-like convection and the penetrative convection; for  $H_s \cong 1$  and  $A \cong 2$ , as in the cases presented in figure 13 for this correlation, a pitchfork bifurcation arises first, due to the unstable layer of fluid at rest above the thermal plume. But this is followed by a resymmetrisation of the stationary regime, the mass of fluid below the source ( $A - H_s \cong 1$ ) being large enough to supply the fluid entrained downstream by the plume and restore the symmetry. For higher Rayleigh numbers, an oscillatory regime occurs through a supercritical Hopf bifurcation indicating that this regime can be classified as the penetrative convection regime. Such a sequence has been shown by Desrayaud & Lauriat (1993, figure 10) for a rectangular vessel of aspect ratio 2. The third correlation,  $f = 0.226 Ra^{0.4}$ , seems to correspond to the natural swaying regime. At such high Rayleigh numbers very fine grids are needed and this can explain the large discrepancies found. Moreover, there are too few numerical points to extract some trends with confidence. Finally, owing to the geometrical parameters chosen by Igarashi ( $H_s = A/2$  and  $A \leq 2$ ), it was impossible to give prominence to the Rayleigh–Bénard-like regime which appears only when  $H_s > 1$ . It is worth noting that this regime can be found in rectangular vessels of aspect ratio 2.5 (full stars in figure 14). For low and moderate Rayleigh numbers, the frequencies of the Rayleigh–Bénard-like regime show that it does not follow the same type of correlation as the one found by Igarashi. The evolution of the frequencies for a given geometrical configuration does not follow a simple power-law, a sharp increase of the frequencies arising close to the bifurcation point when the Rayleigh number increases. For higher Rayleigh numbers, the evolution of the frequencies seems to be of the type determined by Igarashi, i.e.  $Ra^{0.4}$ , but depends on the depth of immersion.

## 5. Conclusions

The modelling of a heated wire as a cylinder of small diameter has been compared with a heat source formulation. For steady flows at low Rayleigh numbers, in the so-called conduction and transition regimes, fairly large discrepancies were found between the velocity fields because of the weak motion. However, the temperature field being unaffected by an obstacle, the agreement is much better for the heat transferred at the bottom and top cold surfaces. On the other hand, the velocity fields are in good agreement at high Rayleigh numbers, except just above the wire since the heat is then convected in a well defined plume. Only small differences in the isothermal patterns were found in the stagnant zone below the heat source. One of the most important parameters is the depth of immersion of the heat source. For unsteady regimes, both models show similar dynamical behaviour with very close frequencies for all of the types of bifurcation points investigated. Therefore, the heat source formulation was used to resolve the supercritical flows around a heated wire.

Swaying motions of confined thermal plumes above a horizontal line heat source inside a vessel with adiabatic side walls, cold top and bottom walls were investigated. The numerical predictions of the swaying frequency of laminar plumes were found in very good agreement with the experimental correlations of Igarashi (1978) and Noto (1989). A variety of dynamic behaviours were shown according to the depth of immersion, aspect ratio and Rayleigh number.

For rectangular vessels with  $A > 1$ , two destabilizing mechanisms characterized by low frequency motions were found:

- for depths of immersion greater than the width of the vessel and small enough Rayleigh numbers, Rayleigh-Bénard-like instabilities may appear within the layer of fluid above the thermal plume. This results in asymmetric steady motions occurring through a pitchfork bifurcation, followed by a subcritical Hopf bifurcation.

- for depths of immersion smaller than the width, the mechanism driving the periodic motion is the penetrative convection which occurs within the layer of fluid at rest below the line heat source.

For square vessels, penetrative convection appears for rather small depths of immersion. Otherwise, for large depths of immersion, steady symmetric flows exist at high Rayleigh numbers. For  $Ra$  greater than a critical value which depends on the depth of immersion, a swaying motion with high frequency starts. This motion is followed by a two-frequency locked regime, then a weakly turbulent regime arises via an intermittent route to chaos.

## List of symbols

$A = H/L$	vertical aspect ratio;
$a$	thermal diffusivity;
$d$	dimensionless cylinder diameter;
$f$	frequency of time-dependent motion;
$g$	gravitational acceleration;
$H$	height of the vessel;
$H_s = A - y_s$	depth of immersion;

$\mathbf{k}$	unit vector in the $y$ -direction;
$L$	width of the vessel;
$p$	pressure;
$Pr = \nu_0/a_0$	Prandtl number;
$Q$	heat generated per unit length (W/m);
$Ra = g\beta QL^3/\lambda_0 a_0 \nu_0$	Rayleigh number;
$t$	time;
$\Delta t$	dimensionless time step;
$T$	temperature;
$T_a$	ambient temperature;
$\mathbf{V}$	velocity vector of components ( $U, V$ )
$x, y$	coordinates;
$\beta$	volumetric thermal expansion coefficient;
$\delta$	Dirac delta function;
$\theta = \lambda_0(T - T_a)/Q$	dimensionless temperature;
$\lambda$	thermal conductivity;
$\mu$	fluid viscosity;
$\nu$	kinematic fluid viscosity;
$\rho$	fluid density;
$\phi$	dimensionless heat flux through a horizontal plane $y$ ;
$\psi$	dimensionless stream function;
$\Omega$	dimensionless vorticity.

### Subscripts

$0$	for reference temperature,
$s$	refers to line source;

### Superscripts

$'$	dimensional quantity.
-----	-----------------------

## References

- Beck J V, McLain H A, Karnitz M A, Shonder J A, Segan E G 1988 Heat losses from underground steam pipes. *J. Heat Transfer* 110: 814–820
- Bill R G, Gebhart B 1975 The transition of the plane plumes. *Int. J. Heat Mass Transfer* 18: 513–526
- Brodowicz K, Kierkus W T 1966 Experimental investigation of laminar free-convection flow in air above horizontal wire with constant heat flux. *Int. J. Heat Mass Transfer* 9: 81–94
- Crane L J 1959 Thermal convection from horizontal wire. *Z. Angew. Math. Phys.* 10: 453–460
- Deschamps V, Desrayaud G 1994 Modeling a horizontal heat-flux cylinder as a line source. *AIAA J. Thermophys. Heat Transfer* 8(1): 84–91
- Deschamps V, Saia J H 1992 Finite element simulation of thermal convection from horizontal cylinders embedded in boxes. *Proc. 3rd European FIDAP Users Group Meeting, Heidelberg*, pp. 1–15
- Desrayaud G, Lauriat G 1991 On the difficulties in computing bifurcation points: application to buoyant plumes. *High performance computing II* (eds) M Durand, F El Dabaghi (Amsterdam: Elsevier Science, North-Holland)
- Desrayaud G, Lauriat G 1993 Unsteady confined buoyant plumes. *J. Fluid Mech.* 252: 617–646
- Desrayaud G, Lepeutrec Y, Lauriat G 1990 Numerical simulation of oscillatory convection in low-Pr fluids. *Notes on numerical fluid mechanics* (Braunschweig: Vieweg) 27: 49–56

- Eichhorn R, Lienhard J H, Chen C C 1974 Natural convection from isothermal spheres and cylinders immersed in a stratified fluid. *Proc. 5th Int. Heat Transfer Conf.*, Tokyo 3: 10–14
- Eichhorn R, Vedhanayagam M 1982 The swaying frequency of line source plumes. *Proc. 7th Int. Heat Transfer Conf.* Munich 2: 407–412
- Engelman M S, Sani L, Gresho P M, Bercovier M 1982 Consistent versus reduced integration penalty methods for incompressible media using several old and new elements. *Int. J. Numer. Methods Fluids* 2: 25–42
- Farouk B, Shayer H 1985 Natural convection around a heated cylinder buried in a saturated porous medium. *Proc. 23th ASME Heat Transfer Conf.* Denver 46: 181–189
- FIDAP 1991 *Theoretical manual, Vol. 1, Version 6.0* (Evanston, IL: Fluid Dynamics International)
- Forstrom R J, Sparrow E M 1967 Experiments on the buoyant plumes above a heated horizontal wire. *Int. J. Heat Mass Transfer* 10: 321–331
- Fujii T 1963 Theory of the steady laminar natural convection above a horizontal line heat source and a point heat source. *Int. J. Heat Mass Transfer* 6: 597–606
- Fujii T, Fujii M, Honda T 1982 Theoretical and experimental studies of the free convection around a long horizontal thin wire in air. *Proc. 7th Int. Heat Transfer Conf.*, Munich 2: 311–316
- Fujii T, Morioka I, Uehara H 1973 Buoyant plume above horizontal line heat source. *Int. J. Heat Mass Transfer* 16: 755–768
- Gebhart B, Jaluria Y, Mahajan R L, Sammakia B 1988 *Buoyancy-induced flows and transport* (New York: Hemisphere) chap. 3
- Gebhart B, Pera L, Schorr A W 1970 Steady laminar natural convection plumes above a horizontal line heat source. *Int. J. Heat Mass Transfer* 13: 161–171
- Golub G H, Meurant G A 1983 *Résolution numérique des grands systèmes linéaires. Collection de la direction des Etudes et Recherches d'Electricité de France* (ed) Eyrolles, (Paris): chap. 3
- Haaland S E, Sparrow E M 1973 Stability of buoyant boundary layers and plumes, taking into account nonparallelism of the basic flows. *J. Heat Transfer* 95C: 295–301
- Hasnaoui M, Bilgen E, Vasseur P 1990 Natural convection above an array of open cavities heated from below. *Numer. Heat Transfer* 18A: 463–482
- Hieber C A, Nash E J 1975 Natural convection above a line heat source: Higher-order effects and stability. *Int. J. Heat Mass Transfer* 18: 1473–1479
- Himasekhar K, Bau H H 1988 Thermal convection around a heat source embedded in a box containing a saturated porous medium. *J. Heat Transfer* 110: 649–654
- Huyakorn P S, Taylor C, Lee R L, Gresho P M 1978 A comparison of various mixed-interpolation finite elements in the velocity-pressure formulation of the Navier–Stokes equations. *Comput. Fluids* 6: 25–35
- Igarashi T 1978 Natural convective oscillatory flow in an enclosed space. Part 2. *Bull. Jpn. Soc. Mech. Eng.* 21: 1022–1030
- Igarashi T, Kada S 1977 Natural convective oscillatory flow in an enclosed space. Part 1. *Heat Transfer Jpn. Res.* 6(1): 19–40
- Incropera F P, Yaghoubi M A 1980 Buoyancy driven flows originating from heated cylinders submerged in a finite water layer. *Int. J. Heat Mass Transfer* 23: 269–278
- Jaluria Y 1982 Thermal plume interaction with vertical surfaces. *Lett. Heat Mass Transfer* 9: 107–117
- Kimura S, Bejan A 1983 Mechanism for transition to turbulence in buoyant plume flow. *Int. J. Heat Mass Transfer* 26: 1515–1532
- Lauriat G, Desrayaud G 1990 Numerical study of oscillatory buoyant plumes above a horizontal line heat source. *Proc. 9th Int. Heat Transfer Conf.*, Jerusalem 4: 171–176
- Lin H T, Cheng W T 1992 Comprehensive correlations for laminar mixed convection line plume and wall plume. *Int. J. Heat Mass Transfer* 35: 2751–2753
- Lyakhov Y N 1970 Experimental investigation of free convection above a heated horizontal wire. *J. Appl. Mech. Tech. Phys.* 11: 355–359
- Maquet J, Gouesbet G, Berlemont A 1992 Numerical simulation of surface tension- and combined buoyancy-driven convection in a liquid layer heated by a hot wire. *Int. J. Heat Mass Transfer* 35: 2695–2703
- Morwald K, Mitsotakis K, Schneider W 1986 Higher-order analysis of laminar plumes. *Proc. 8th Int. Heat Transfer Conf.*, San Francisco 3: 1335–1340
- Nawoj H J, Hickman R S 1977 An experimental investigation of the plume velocity field above a horizontal line heat source. *J. Heat Transfer* 99: 609–613



- Noto K 1989 Swaying motion in thermal plume above a horizontal line heat source. *J. Thermophys.* 3: 428–434
- Noto K, Matsui S, Matsumoto R 1982 Observation on vortex pair of plane plume in thermally stratified fluid. *Flow visualization* (Berlin: Springer Verlag) vol. 4
- Noto K, Matsumoto R 1986 Swaying motion in buoyant air plume above a horizontal line heat source. *AIAA/ASME 4th Joint Thermophysics and Heat Transfer Conf.*, Boston, paper 86–1266, pp. 1–9
- Pera L, Gebhart B 1971 On the stability of laminar plumes: Some numerical solutions and experiments. *Int. J. Heat Mass Transfer* 14: 975–984
- Pera L, Gebhart B 1975 Laminar plume interactions. *J. Fluid Mech.* 68: 259–271
- Peyret R 1990 The Chebyshev multidomain approach to stiff problems in fluid mechanics. *Methods Appl. Mech. Eng.* 80: 129–145
- Roache P 1982 *Computational fluid dynamics* (Albuquerque: Hermosa)
- Roux B (ed.) 1990 Numerical simulation of oscillatory convection in low-Prandtl fluids. *Notes on numerical fluid mechanics* (Braunschweig: Vieweg) vol. 27
- Reimann J 1974 Experimental investigation of free convection flow from wires in the vicinity of phase interfaces. *Int. J. Heat Mass Transfer* 17: 1051–1061
- Roze C, Gouesbet G, Darrigo R 1993 Oscillatory instabilities produced by heat from a temperature-controlled hot wire below an interface. *J. Fluid Mech.* 250: 253–276
- Schorr A W, Gebhart B 1970 An experimental investigation of natural convection wakes above a line heat source. *Int. J. Heat Mass Transfer* 13: 557–571
- Sohn J L 1988 Evaluation of FIDAP on some classical laminar and turbulent benchmarks. *Int. J. Numer. Methods Fluids* 8: 1469–1490
- Urakawa K, Morioka I, Kiyota M 1983 Swaying motion of the buoyant plume above a horizontal line heat source. *Proc. 1st ASME–JSME Thermal Eng. Conf.*, Honolulu 3: 215–220
- Wakitani S 1985 Non-parallel-flow stability of a two-dimensional buoyant plume. *J. Fluid Mech* 159: 241–258
- Wakitani S, Yosinobu H 1984 Stability characteristics of a natural convection flow above a horizontal line heat source. *J. Phys. Soc. Jpn.* 53: 1291–1300
- Wolf A, Swift J B, Swinney H L, Vastano J A 1985 Determining Lyapounov exponents from a time series. *Physica D* 16: 285–317
- Xia J L, Xin M D, Zhang H J 1990 Natural convection in an externally heated enclosure containing a local heat source. *J. Thermophys. Heat Transfer* 4: 233–238
- Xia J L, Xin M D, Zhang H J 1991 A numerical study of natural convection in an externally heated enclosure with a local heat source. *Proc. 9th Int. Heat Transfer Conf.*, Jerusalem 4: 217–222
- Yaghoubi M A, Incropera F P 1978 Natural convection from a heated horizontal cylinder submerged in a shallow water layer. *Proc. 6th Int. Heat Transfer Conf.*, Toronto 2: 269–274
- Yang H Q 1992 Buckling of a thermal plume. *Int. J. Heat Mass Transfer* 35: 1527–1533
- Yosinobu H, Onishi Y, Amano S, Enyo S, Wakitani S, 1979 Experimental study on instability of a natural convection flow above a horizontal line heat source. *J. Phys. Soc. Jpn.* 47: 312–319



## Interaction between a buoyancy-driven flow and an array of annular cavities

M MOLKI<sup>1</sup> and M FAGHRI<sup>2</sup>

<sup>1</sup>Department of Mechanical Engineering, Esfahan University of Technology, Esfahan, Iran

<sup>2</sup>Department of Mechanical Engineering and Applied Mechanics, University of Rhode Island, Kingston, RI 02881, USA

**Abstract.** A numerical study is performed to investigate the interaction between a buoyancy-induced flow and an array of annular cavities. The buoyant flow is generated in a vertical annular enclosure with a centrally-positioned finned inner cylinder. Heat is generated within the inner cylinder, and it is convected through the inter-fin cavities and annular enclosure to the outside environment. The results indicate the presence of a twin recirculating bubble in each cavity. At higher  $Ra$ , the main flow enters the cavities and removes the recirculating flow. These observations are more pronounced at higher  $Pr$ . For more slender and deeper cavities, the recirculating bubbles closer to the finned wall collapse and split into two bubbles. The presence of cavities create a nearly uniform high-temperature zone adjacent to the finned wall. As the fin length is reduced and the cavities become more shallow, this zone shrinks and the main buoyancy-driven flow maintains a closer thermal communication with the finned wall.

**Keywords.** Heat transfer; cavity; buoyancy; annular.

### 1. Introduction

Buoyancy-driven flows in enclosures have been the subject of study for many years and the published literature on the subject is diverse and rich (Catton 1978; Ostrach 1982, 1988; Hoogendoorn 1986). A number of investigators have been interested in cavity-driven flow with the flow generated by the motion of the upper wall (Bozeman & Dalton 1973; Ozawa 1975; Tuann & Olson 1978; Benjamin & Denny 1979; Ghia *et al* 1982; Schreiber & Keller 1983; Iwatsu *et al* 1989, 1990, 1992) while others have studied the effect of thermally-induced buoyancy on the flow field.

Although the existing literature on square cavities is rich, the number of investigations concerned with annular cavities is fairly limited. The focus of the previous studies has been in two directions; one to gain better understanding of the flow field, and the other to explore the thermal behaviour of the problem. In fact, the coupling of the governing equations in thermally-driven flows links these objectives together.

One of the pioneering works on natural convection in annular cavities was performed by De Vahl Davis & Thomas (1969) who employed a numerical technique and reported flow and heat transfer results. In another effort, Thomas & De Vahl Davis (1970) classified the flow into three regimes, namely conduction, transition, and boundary layer regimes, and presented correlations for each regime. Other related works are the experimental investigations of Collier *et al* (1970), who employed carbon dioxide at high pressures and temperatures, the analytical and experimental studies of Nagendra *et al* (1970), and the experimental work of Sheriff (1966).

Experiments on a vertical annulus where the inner cylinder is at constant heat flux and the outer cylinder at constant temperature are performed by Keyhani *et al* (1983). They used air and helium as working fluids and conducted their measurements at a fixed aspect and radius ratio. Bhushan *et al* (1983) continued the studies of Keyhani *et al* (1983) with the same experimental setup and performed experiments with air and helium at higher pressures. They arrived at new and more general correlations for conduction and boundary layer regimes which included the effect of aspect and radius ratio on heat transfer coefficient. However, the results are valid for the annuli with constant surface heat flux at the inner cylinder.

Experiments of Prasad & Kulacki (1985) were performed at higher values of Rayleigh number and with the inner and outer surfaces maintained at constant temperature. They indicated that turbulence is initiated when the local Grashof number reaches  $4 \times 10^9$ . Other related studies are the experiments reported by Vijayan *et al* (1986) on two methods of reduction of buoyancy-driven flow in a vertical annulus open to hot fluid at the bottom, the numerical results of Lin & Nansteel (1987) for water at maximum density, the experimental work of Molki & Shahsavan (1989) on a vertical annulus filled with atmospheric air and immersed in a water bath to provide a convective environment on the outside surface of the annulus, and the numerical results reported by Farouk *et al* (1990).

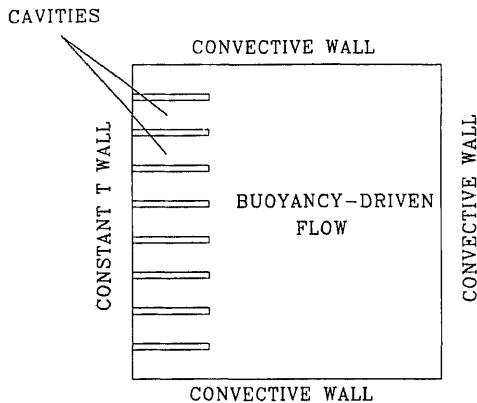
Despite the existence of the above references on buoyancy-driven flows in annularities, the interaction between annular cavities appears to be scarce. The present investigation is concerned with the flow and temperature fields which result from the interaction between a buoyancy-induced flow and an array of annular cavities. The flow is induced by thermal gradient in an annular enclosure where a finned inner cylinder is situated at its centre. The interfin spacings form the array of open cavities which are exposed to the buoyant flow.

This type of interaction arises in many engineering devices. For instance, the finned surfaces form an array of open cavities whose thermal-hydraulic performance depends on the nature of its interaction with the neighbouring fluid. In particular, the inner cylinder of the present geometry may be considered to be a heat-generating electronic component which is shrouded in a cylindrical container to protect it against the harsh environment. This aspect of the buoyancy-driven flow appears to be new and has not attracted the attention of other investigators.

## 2. Description, formulation, and methodology

### 2.1 Statement of the problem

The simplified view of the problem is shown in figure 1. A solid cylinder of radius  $R_i$  and height  $H$  is located in a larger hollow cylinder of radius  $R_o$  to form the annular enclosure. In this figure only the enclosure and the finned surface of the inner cylinder



**Figure 1.** Geometry of the problem; the finned surface at the left is the surface of the inner cylinder.

are shown. The fins are equally spaced to form the array of annular cavities. Heat is generated within the inner cylinder at the uniform rate  $s$  per unit volume. The outer boundaries of the enclosure are exposed to a convective environment at  $T_\infty$  and  $h$ . With this arrangement, the thermal energy generated in the finned wall is transferred to the outside environment by natural convection through the annular fluid. The circumferential fins form the array of open cavities and interact with the buoyancy-driven main flow in the annulus. It should be noted that the solutions are based on the large value of thermal conductivity for the inner wall so that the thermal boundary condition on the finned wall may be considered to be uniform temperature.

## 2.2 Governing differential equations

The governing equations are conservation of mass, momentum, and energy for the fluid, and the energy equation for the inner finned solid. All the thermophysical properties are assumed to be constant, except for the density in the buoyancy term where Boussinesq approximation is used. The dimensionless governing differential equations are,

$$\frac{1}{R} \frac{\partial}{\partial R} (RV) + \frac{\partial U}{\partial X} = 0, \quad (1)$$

$$U \frac{\partial U}{\partial X} + V \frac{\partial U}{\partial R} = -\frac{\partial P}{\partial X} + \left[ \frac{1}{R} \frac{\partial}{\partial R} \left( R \frac{\partial U}{\partial R} \right) + \frac{\partial^2 U}{\partial X^2} \right] + \frac{Ra}{Pr} T, \quad (2)$$

$$U \frac{\partial V}{\partial X} + V \frac{\partial V}{\partial R} = -\frac{\partial P}{\partial R} + \left[ \frac{1}{R} \frac{\partial}{\partial R} \left( R \frac{\partial V}{\partial R} \right) + \frac{\partial^2 V}{\partial X^2} \right] - \frac{V}{R^2}, \quad (3)$$

$$U \frac{\partial T}{\partial X} + V \frac{\partial T}{\partial R} = \frac{1}{Pr} \left[ \frac{1}{R} \frac{\partial}{\partial R} \left( R \frac{\partial T}{\partial R} \right) + \frac{\partial^2 T}{\partial X^2} \right] + \frac{KS}{Pr}, \quad (4)$$

$$\frac{1}{R} \frac{\partial}{\partial R} \left( R \frac{\partial T}{\partial R} \right) + \frac{\partial^2 T}{\partial X^2} + S = 0. \quad (5)$$

In these equations, the dimensionless parameters are defined as,

$$X = X/L, \quad R = r/L, \quad (6)$$

$$U = u/(v/L), \quad V = v/(v/L), \quad T = t/t_\infty, \quad (7)$$

$$P = p^*/[\rho(v/L)^2], \quad p^* = p + \rho_r g(1 + \beta T_r)(H - x), \quad (8)$$

$$Ra = g\beta L^3 T_\infty/(\nu\alpha), \quad Pr = (\nu/\alpha), \quad K = (k_s/k_f), \quad S = (L^2/k_s T_\infty)s. \quad (9)$$

To complete the formulation of the problem, the boundary conditions are specified as: (1) on all solid boundaries,  $U = V = 0$ ; (2) at the centreline of the inner cylinder (not shown in figure 1),

$$\partial T/\partial R = 0; \quad (10)$$

(3) at the top and bottom walls,

$$\text{solid} \quad \pm \partial T/\partial X = Bi(T - 1), \quad (11)$$

$$\text{fluid} \quad \mp \partial T/\partial X = Nu_\infty(T - 1), \quad (12)$$

where the negative and positive signs apply respectively to the top and bottom walls; (4) at the outer cylindrical wall,

$$-\partial T/\partial R = Nu_\infty(T - 1), \quad (13)$$

(5) at the solid-fluid interface,  $T_s = T_f$  and

$$K(\partial T/\partial N)_s = (\partial T/\partial N)_f. \quad (14)$$

In these equations,  $Bi = hL/k_s$ ,  $Nu_\infty = hL/k_f$ , and  $N = n/L$ , where  $n$  is a space coordinate normal to the interface. Moreover, the geometric parameters which identify the annulus and the fins are, radius of the outer cylinder  $R_o = r_o/L$ , annulus height  $H = h/L$ , fin length  $L_f = l_f/L$ , fin thickness  $T_f = t_f/L$ , and the number of fins  $n_f$ . It is noteworthy that the radius of the inner cylinder  $R_i = r_i/L$ , is related to  $R_o$  through  $R_i = R_o - 1$ . Once one of the two radii is known, the other is readily found. In this connection, the ratio of the outer radius to the inner radius,  $r_o/r_i = R_o/R_i$ , has emerged as a major geometric parameter.

Examination of (1)–(9) and the boundary conditions indicate that the dependent variables, namely  $U$ ,  $V$ ,  $P$ , and  $T$ , are functions of  $X$ ,  $R$ ,  $Ra$ ,  $Pr$ ,  $K$ ,  $Nu_\infty$ ,  $S$ ,  $R_o/R_i$ ,  $H$ ,  $L_f$ ,  $T_f$ , and  $n_f$ . In this study,  $Ra$  ranged from 1000 to 1000,000,  $Pr$  from 0.7 to 1000,  $L_f$  from 1/64 to 1/2,  $T_f$  from 0.009 to 0.087, and  $n_f$  from 0 to 16. To save on the computation time, the remaining variables were fixed at values  $K = 2800$ ,  $Nu_\infty = 20$ ,  $S = 0.04$ ,  $R_o/R_i = 2$ , and  $H = 1$ . This value of  $K$  corresponds to a practical situation where the inner cylinder is made of iron and the cavity fluid is air. The values for  $Nu_\infty$  and  $S$  are other realistic values close to those encountered in the studies of Molki & Shahsavan (1989).

### 2.3 Computational method

The governing equations were discretized by a control-volume-based finite-difference method and solved by a line-by-line iterative procedure. In this connection, the momentum and energy equations were integrated over each control volume of the solution domain. Temperatures were evaluated at the centre of main control volumes, while a staggered grid was employed for velocity components. The continuity

equation was integrated on the main control volumes to yield a pressure-correction equation.

Before integrating the differential equations, the convection and diffusion terms were combined to form a total flux term. A power-law scheme was used to interpolate total fluxes. The interface diffusion coefficients were obtained from the harmonic mean formula. The source terms were linearized and the convective boundary conditions were supplied as additional source terms. Care was exercised to avoid negative-slope linearization of the source terms, which could otherwise result in a diverged solution. Due to nonlinearity of the differential equations, underrelaxation had to be applied to the velocity components and the pressure. The underrelaxation values 0.5 and 0.8, respectively for velocities and pressure, proved to be quite satisfactory. More details of the discretization and computational method are well documented by Patankar (1980).

Before the onset of computational runs, a comprehensive grid study was performed on the problem. The total number of grids were 1160, 2350, 3900, 4290, and 7968, while the various parameters were kept constant at  $Ra = 10,000$ ,  $Pr = 0.7$ ,  $Nu_\infty = 20$ ,  $S = 0.04$ ,  $K = 2800$ ,  $R_o/R_i = 2$  and  $H = 1$ . The effect of grid size on  $U$ ,  $V$ , and  $T$  at the centre of the cavity indicated that these parameters changed with respect to the values for the finest grid (i.e., the 7968-point mesh) by (5.32%, 1.78%, 0.61%, 0.61%), (1.07%, 0.61%, 0.36%, 0.36%), and (0.19%, 0.12%, 0.06%, 0.07%), respectively. These percentages indicated that a total of 3900 grid points would be sufficient to perform the present computations.

In the iterative procedure, the value of mass source in control volumes served as a convergence criteria. The iterations were continued until the maximum value of mass source did not exceed  $10^{-7}$ . With this criteria, the number of iterations ranged from 600 to 1200. All computations were performed in double precision on a VAX 4000/200 computer. The CPU time ranged from 1 to 2 hours, with the 2-hour runs corresponding to the high Rayleigh numbers.

### 3. Comparison with bench mark solution

The precision of the present numerical technique and the computer code was evaluated by comparison with a bench mark solution. Apparently, no such solution is available for annular cavities. Therefore, the comparison is made with the rectangular cavity. It should be noted, however, that the rectangular cavity may be considered as a special case of the annular geometry where the radius ratio is equal to one. In fact, this comparison could also serve as a test of performance for the present code in such an extreme limiting case.

To facilitate the comparison with the bench mark solution of De Vahl Davis (1983), the radii of the cavity were increased to allow the radial inter-wall distance between the vertical walls of the cavity to reach 2.4% of the outer radius. This also increased the extent of the solution domain and the computation required more grid points. In addition, the top and bottom walls of the cavity were insulated, and the values of  $Nu_\infty$  and  $K$  were increased respectively to  $2 \times 10^8$  and  $2.8 \times 10^8$  to ensure a uniform temperature on the vertical walls of the enclosure.

In this comparison, the computed quantities are the maximum vertical velocity on the horizontal mid-plane of the cavity, the maximum horizontal velocity on the vertical mid-plane of the cavity, the respective locations of these velocities, and the

Nusselt number. The numerical results indicated a maximum difference of 2.0% in velocities and 1.7% in Nusselt number. This level of agreement between the present work and the bench mark solution under such an extreme limiting case is supportive of the present computational technique. Additional comparison with literature, including a comparison with one of the previous experimental results of the authors (Molki & Shahsavan 1989) has been presented by Molki & Faghri (1994).

**Table 1.** Values of parameters for figures 2–6.

Figure	2a	2b	2c	2d	
Pr	0.7	0.7	1000	1000	
Ra	1000	1000,000	1000	1000,000	
$n_f$	8	8	8	8	
$L_f$	0.25	0.25	0.25	0.25	
$T_f$	0.02	0.02	0.02	0.02	
$S_f$	0.093	0.093	0.093	0.093	
	3a	3b	3c	3d	3e
	0.7	0.7	0.7	0.7	0.7
	100,000	100,000	100,000	100,000	100,000
	8	8	8	8	8
	0.25	0.125	0.063	0.031	0.016
	0.02	0.02	0.02	0.02	0.02
	0.093	0.093	0.093	0.093	0.093
	4a	4b	4c	4d	
	0.7	0.7	0.7	0.7	
	100,000	100,000	100,000	100,000	
	12	6	4	2	
	0.25	0.25	0.25	0.25	
	0.02	0.02	0.02	0.02	
	0.0585	0.1257	0.184	0.320	
	5a	5b	5c	5d	
	0.7	0.7	0.7	0.7	
	100,000	100,000	100,000	100,000	
	8	8	8	8	
	0.5	0.25	0.125	0.063	
	0.009	0.02	0.042	0.087	
	0.1031	0.0933	0.0735	0.0336	
	6a	6b	6c	6d	
	0.7	0.7	0.7	0.7	
	100,000	100,000	100,000	100,000	
	16	8	4	2	
	0.25	0.25	0.25	0.25	
	0.01	0.02	0.04	0.08	
	0.049	0.093	0.168	0.280	



#### 4. Results and discussion

Attention is now turned to the examination of the results. The results are grouped into two parts. Part one deals with the flow field, while part two considers the temperature. The principal parameters of each case are shown in tables 1 and 2.

**Table 2.** Values of parameters for figures 7–11.

Figure	7a	7b			
Pr	0.7	1000			
Ra	1000	1000,000			
$n_f$	8	8			
$L_f$	0.25	0.25			
$T_f$	0.02	0.02			
$S_f$	0.093	0.093			
	8a	8b	8c	8d	8e
	0.7	0.7	0.7	0.7	0.7
	100,000	100,000	100,000	100,000	100,000
	8	8	8	8	8
	0.25	0.125	0.063	0.031	0.016
	0.02	0.02	0.02	0.02	0.02
	0.093	0.093	0.093	0.093	0.093
	9a	9b			
	0.7	0.7			
	100,000	100,000			
	12	2			
	0.25	0.25			
	0.02	0.02			
	0.0585	0.32			
	10a	10b	10c	10d	
	0.7	0.7	0.7	0.7	
	100,000	100,000	100,000	100,000	
	8	8	8	8	
	0.5	0.25	0.125	0.063	
	0.009	0.02	0.042	0.087	
	0.103	0.093	0.0735	0.0336	
	11a	11b	11c	11d	
	0.7	0.7	0.7	0.7	
	100,000	100,000	100,000	100,000	
	16	8	4	2	
	0.25	0.25	0.25	0.25	
	0.01	0.02	0.04	0.08	
	0.0494	0.093	0.168	0.280	

#### 4.1 The flow field

The interaction between the main buoyancy-driven flow and the array of annular cavities is demonstrated by the streamlines of figure 2. In this figure, the right boundaries of the inter-fin cavities are exposed to the large-scale recirculation in the main enclosure (see figure 1), while the top, bottom, and left sides are bounded by the solid walls. The wall and the fins are warmer than the neighbouring fluid and generate a clockwise rotation in the main buoyancy-driven flow (figure 1).

It is seen from table 1 that figures 2a & b and figures 2c & d are prepared for the same Prandtl number ( $Pr$ ) but different Rayleigh numbers ( $Ra$ ). Examination of these streamlines indicates that at low Rayleigh numbers (figure 2a & c), the cavities are washed by twin recirculating flows which rotate in opposite directions. The recirculating bubbles situated to the right have a counter clockwise rotation, while those located to the left have a clockwise rotation. The top and bottom cavities, however, are somewhat different, since their exposure to the main large-scale recirculation is different. In these specific locations, only one single bubble is observed, which is larger and more energetic.

At higher values of  $Ra$  (figures 2b & d), the main flow enters the cavities of the array where it partially disturbs or completely removes the twin recirculating bubbles. The streamlines in these figures indicate that the extent of this interaction increases with  $Pr$ , and the main recirculating flow moves into the cavities. It is noteworthy

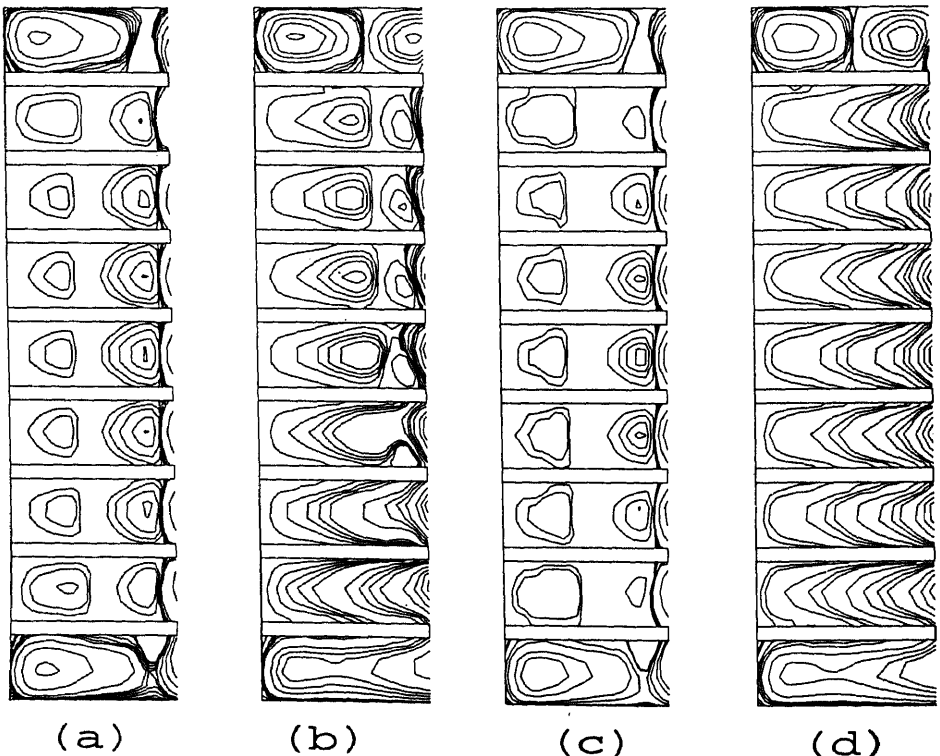
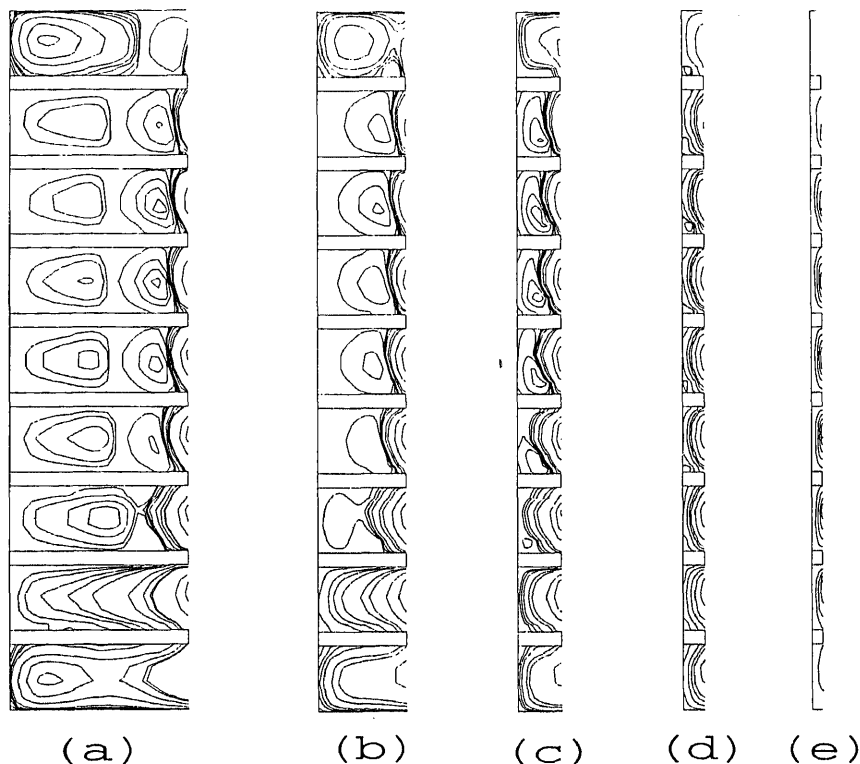


Figure 2. Effect of  $Pr$  and  $Ra$  on the flow field in the array of cavities; twin recirculating bubbles are seen inside most of the cavities.



**Figure 3.** Effect of fin length on the flow field. The recirculating bubbles become smaller and eventually disappear as the cavities become more shallow.

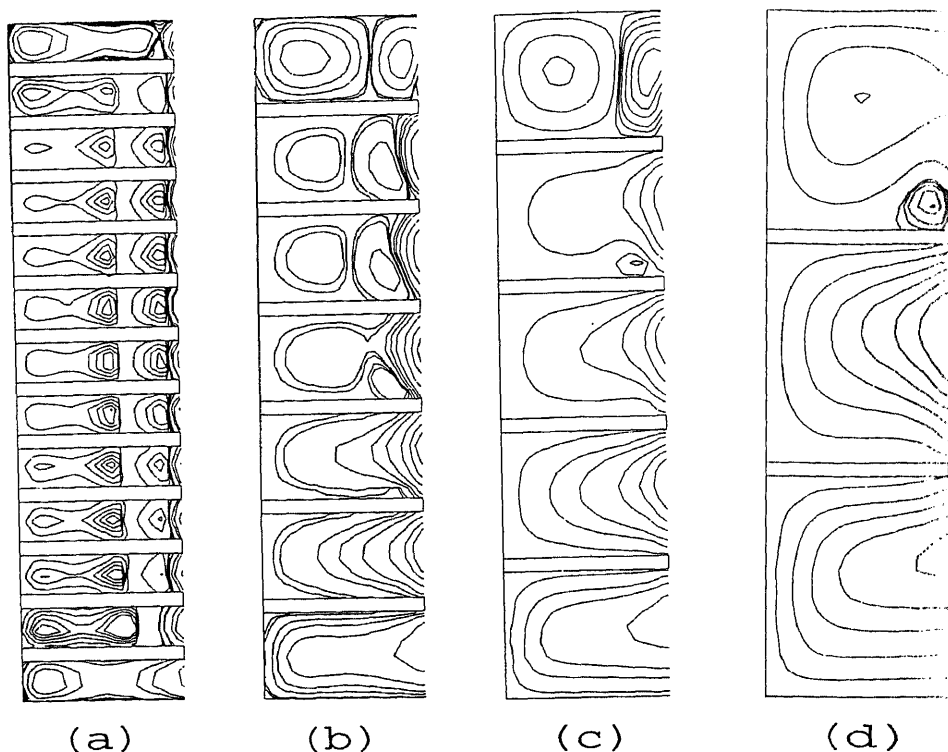
that these flow features are often encountered in finned walls, and they can have a significant effect on heat transfer.

Effect of fin length or depth of cavities on the flow field is shown in figure 3, and the relevant parameters are given in table 1. It is seen from the figure that the twin recirculating bubbles are greatly affected by the fin length. The bubble adjacent to the wall disappears as the fin length is reduced and the cavities become shallow (figure 3a). Further reduction of the fin length brings the main recirculating flow closer to the cavities and there is more interaction with the cavity fluid. It is clear from figures 3d–e that for shallow cavities, the recirculating bubbles disappear and the main flow enters the inter-fin spacings with no difficulty.

In figure 4, the streamlines are presented for different number of fins. The cavities become more slender as the number of fins is increased. In this case, the next-to-the-wall recirculating bubbles tend to collapse and split into two parts (figure 4a). On the other hand, the extent of flow interaction increases as the number of fins is decreased and the main flow moves into the interfin spacings (figure 4d).

The finned walls are often optimized with respect to their thickness and number. In this regard, the fin length can be either increased with constant thickness, or it can be increased under the constraint that the fin mass remains constant. With this constraint, a longer fin has to be thinner and a shorter fin must be thicker.

The streamlines of figures 5–6 are obtained with the restriction that the total fin



**Figure 4.** Effect of number of fins on the flow field. Larger number of fins tends to stretch the next-to-the-wall recirculating bubbles and split them into two parts.

mass remains constant. In figure 5, the fin length is reduced under the constant-mass constraint and the number of fins is constant, while in figure 6, the fin length is held constant and the number of fins is reduced. It is clear from figure 5 that the number and extent of the recirculating bubbles are reduced with the fin thickness. However, as evidenced from figure 6, the main flow moves into the cavities as the number of fins is decreased.

#### 4.2 The temperature field

Representative isotherms for the array of cavities are presented in figures 7–11, with the respective parameters given in table 2. In figure 7, the effect of  $Pr$  and  $Ra$  on temperature distribution indicates that there is a larger temperature variation in the top and bottom cavities of the array, while the temperature gradient in the remaining cavities is relatively small. This observation is consistent with the fact that the flow in the top and bottom cavities is more dynamic. As is evident from the earlier streamlines, the main flow often penetrates into these cavities with less difficulty, thus permitting better exchange of thermal energy.

Figure 8 shows the effect of fin length on the temperature field. According to table 2, the figure is prepared for a fixed  $Pr$  and  $Ra$ , with the fin thickness being constant. Again, it is seen that the temperature variation in the cavities situated at the far ends

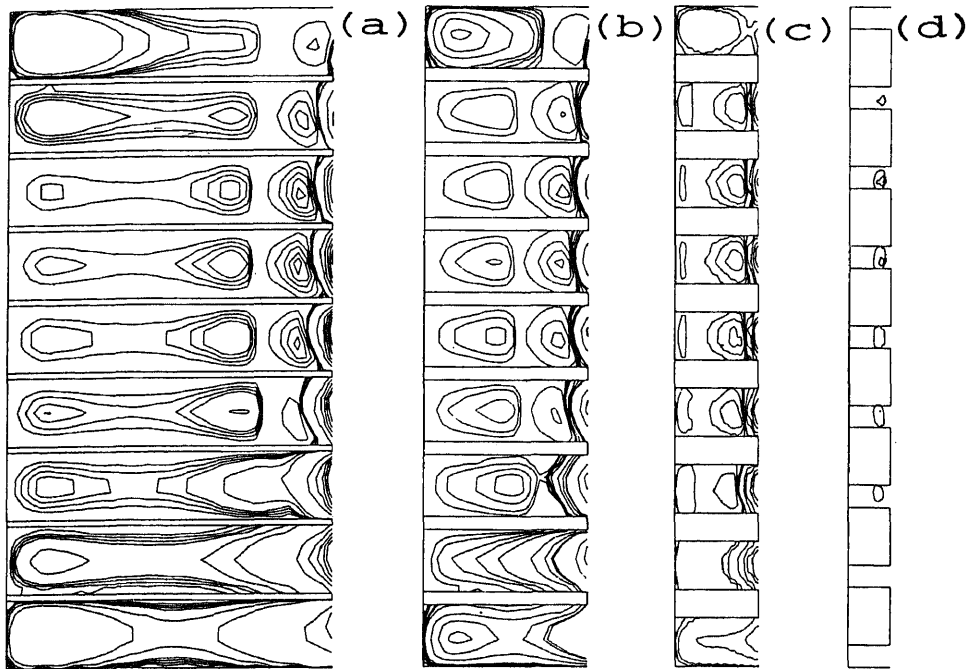


Figure 5. Effect of fin length under constant-mass restriction; the number and extent of the bubbles are reduced with the thickness.

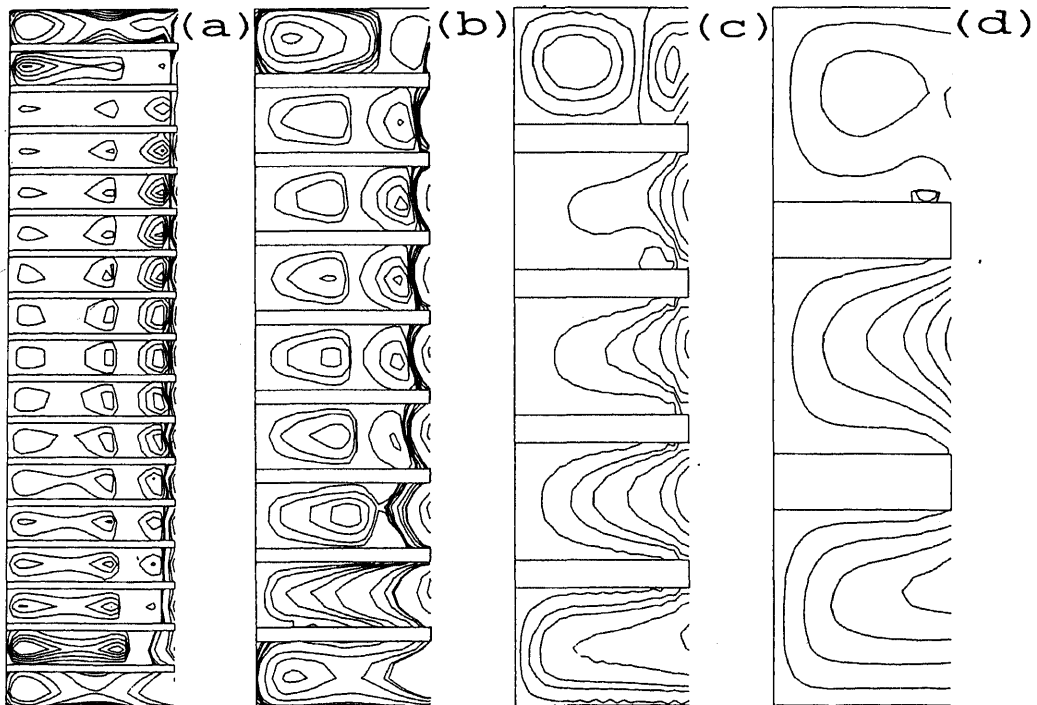


Figure 6. Effect of number of fins under constant-mass restriction; the main flow enters the cavities as the number of fins is decreased.

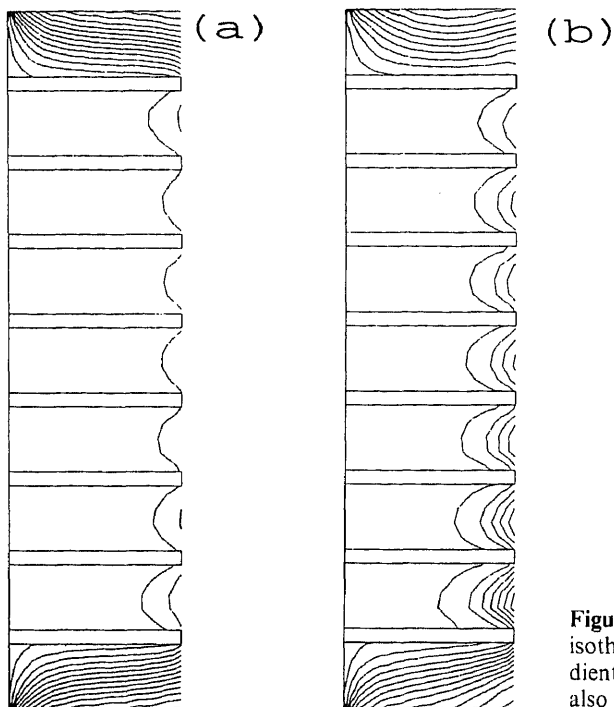
d)

As the number of fins tends to parts.

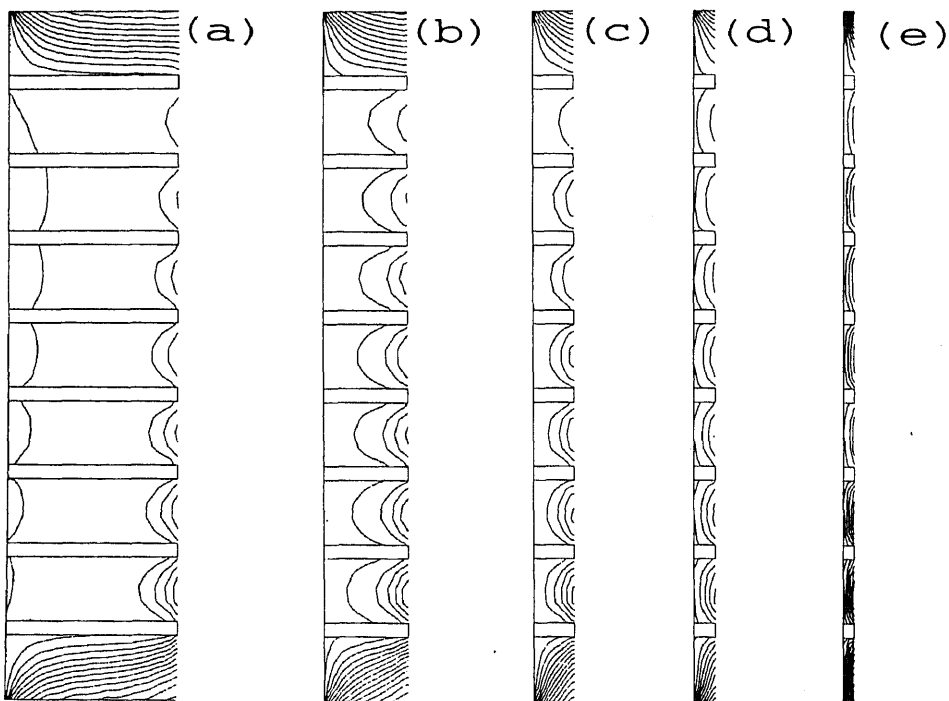
Under constant-mass restriction, the number of bubbles is held constant. However, the number of

As the number of fins is held constant, the number of bubbles in the remaining flow is earlier, thus

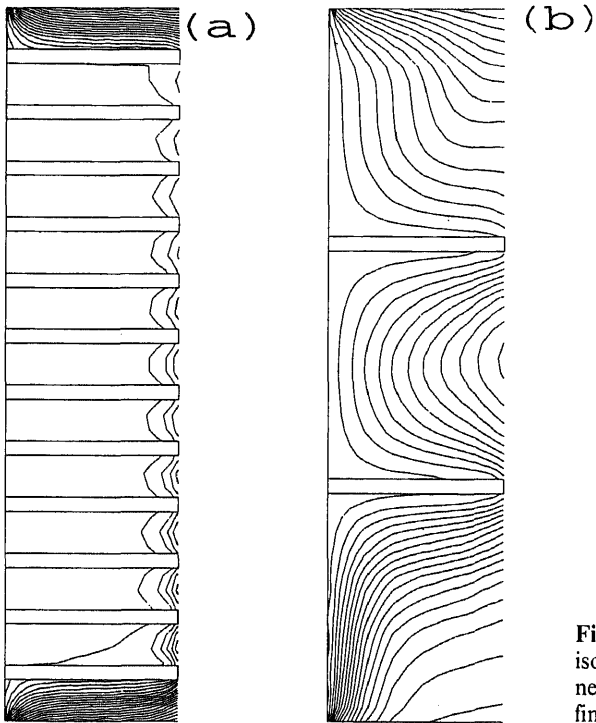
As the number of fins is held constant, the number of bubbles far ends



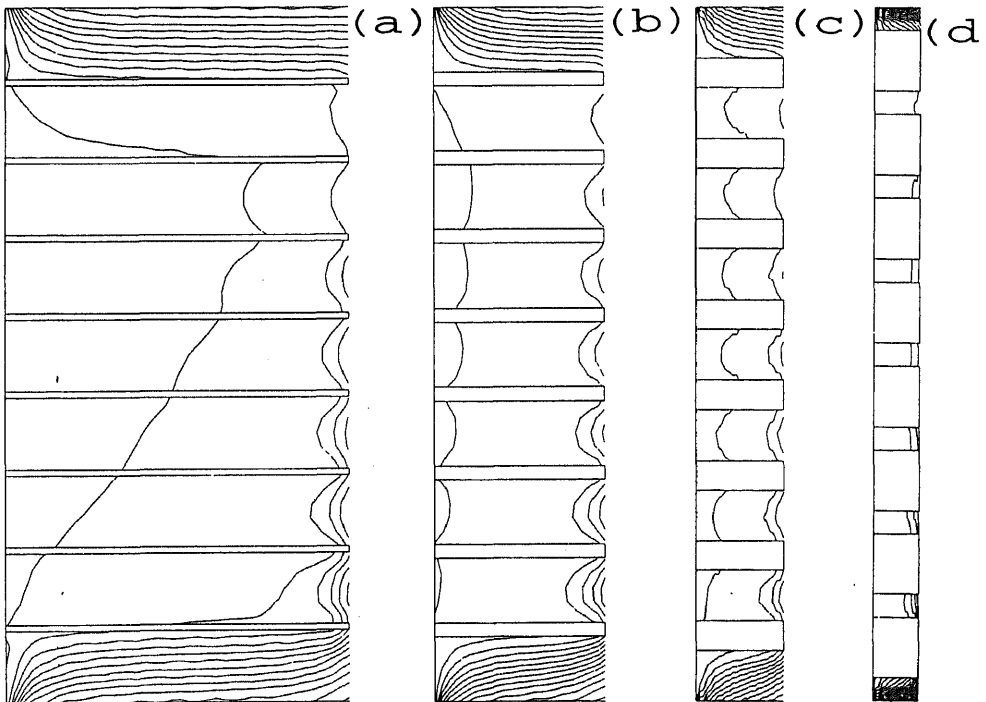
**Figure 7.** Effect of  $Pr$  and  $Ra$  on isotherms; larger temperature gradients are seen near the fin tips and also in the top and bottom cavities.



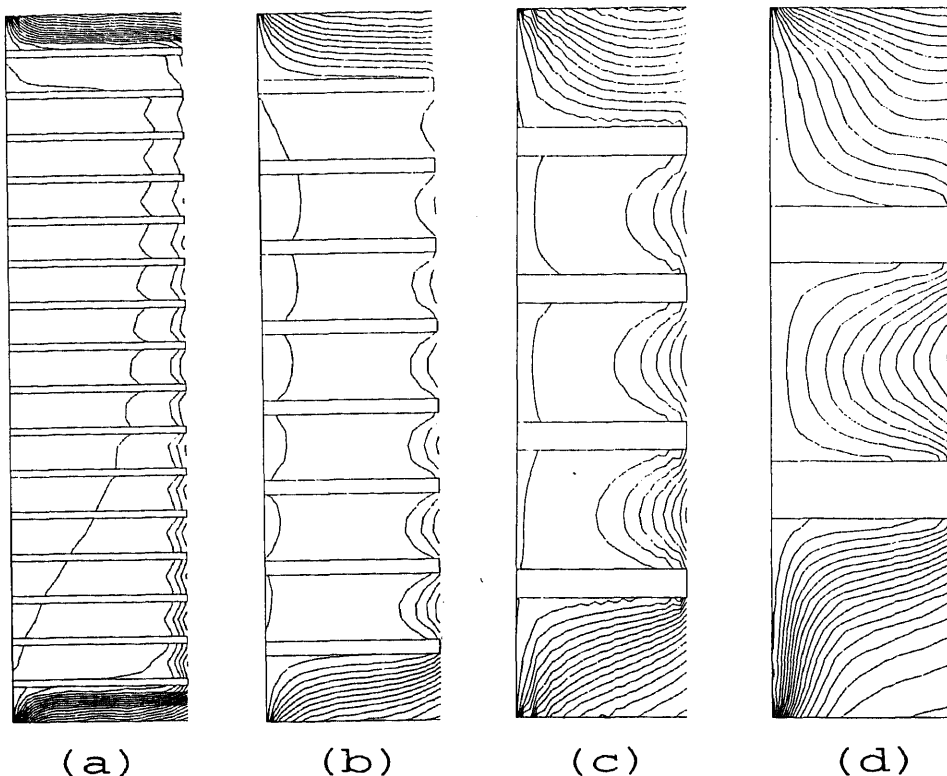
**Figure 8.** Effect of fin length on isotherms; more temperature variation is observed as the fin length is reduced.



**Figure 9.** Effect of number of fins on isotherms; for larger number of fins, a nearly isothermal zone is seen in the finned area.



**Figure 10.** Effect of fin length on isotherms under constant-mass restriction.



**Figure 11.** Effect of number of fins on isotherms under constant-mass restriction.

of the array is much larger. As the fin length is reduced, the extent of the nearly uniform temperature zone created by the presence of the fins is also reduced and there is a closer thermal communication between the finned wall and the main buoyancy-driven flow.

Figure 9 indicates the effect of number of fins on the cavity isotherms. With more fins, the recirculating bubbles are present in the cavities and the thermal communication between the finned wall and the main flow is through the multiple bubbles. As the number of fins is diminished, the size of cavities in vertical direction is increased and the main fluid comes into a close contact with the finned wall. It is seen in this figure that the temperature variation increases with the fin spacing.

At a given temperature difference, there are two parameters which determine the magnitude of heat transfer. These are the heat transfer coefficient and the surface area. As the fin length or the number of fins is reduced, the area decreases and this tends to decrease heat transfer from the wall. On the other hand, the presence of recirculation zones in the inter-fin spacings has an enhancing effect on heat transfer coefficient, so that the overall effect is determined by the combined effect of these parameters. In a previous study, the authors (Molki & Faghri 1994) have indicated that for the case of constant heat generation within the finned wall, the wall temperature decreases as the number of fins or their length is increased.

Isotherms for fin arrangements of constant mass are shown in figures 10–11. Again, the temperature variation in the cavities situated at the far ends of the array is



relatively large. The figures also indicate that the effect of fin spacing (or number of fins) on isotherms is stronger than that caused by the fin length.

## 5. Conclusion

The paper described a numerical investigation of the effect of a buoyancy-driven flow on an array of annular cavities. The buoyant flow is generated by temperature gradient in an annular enclosure where the resulting flow interacts with an array of smaller cavities inside the enclosure. The study is focused on the qualitative description of the flow and temperature fields.

At low Ra numbers, a twin recirculating bubble is observed in the array of cavities which rotate in opposite directions. As Ra increases, flow patterns undergo a transition so that at higher Ra, the main flow is able to enter the cavities and remove the recirculation bubbles. These observations are more pronounced at higher Pr number.

Another noteworthy parameter which affects the flow patterns is the depth of cavities or fin length. As the cavity depth is reduced, the recirculation bubbles adjacent to the bottom wall of cavities disappear. Further decrease of depth would remove all recirculation from within the cavities and would bring the main flow into a closer thermal communication with the finned wall.

Flow patterns are also affected by the vertical dimension of cavities. For more slender and deeper cavities, the recirculating bubbles that are closer to the finned wall collapse and split into two bubbles.

Examination of the temperature field indicates that there is a large temperature variation at the top and bottom cavities of the array, while the temperature gradient in the remaining cavities is relatively small. An overall review of the isotherms shows that the presence of the cavities creates a nearly uniform high-temperature zone adjacent to the finned wall. However, as the fin length is reduced and the cavities become more shallow, this uniform-temperature zone is shrunk and the main buoyancy-driven flow maintains a closer thermal communication with the finned wall.

The first author gratefully acknowledges that the funding for this project was provided by the Esfahan University of Technology.

## List of symbols

Bi	Biot number, $hL/k_s$ ;
$g$	acceleration of gravity, $\text{m/s}^2$ , (8);
$h$	outside convective heat transfer coefficient, $\text{W/m}^2\text{ }^\circ\text{C}$ , annulus height, m;
$H$	dimensionless annulus height, $h/L$ ;
$K$	solid-fluid conductivity ratio, $k_s/k_f$ ;
$l_f$	fin length, m;
$L_f$	dimensionless fin length, $l_f/L$ ;
$L$	radial distance between the inner and outer walls of the annulus, m;
$n$	space coordinate normal to solid-liquid interface;

$n_f$	number of fins;
$N$	dimensionless $n$ , $n/L$ ;
$Nu_\infty$	Nusselt number outside the annulus, $hL/k_f$ , (12);
$p$	pressure, Pa, (8);
$P$	dimensionless pressure, (8);
$Pr$	Prandtl number, $\nu/\alpha$ ;
$R$	dimensionless radial coordinate, $r/L$ , (6);
$r_i$	radius of the inner cylinder, m;
$r_o$	radius of the outer cylinder, m;
$R_i$	dimensionless radius of the inner cylinder, $r_i/L$ ;
$R_o$	dimensionless radius of the outer cylinder, $r_o/L$ ;
$Ra$	Rayleigh number, (9);
$s$	heat generation within the inner cylinder, $W/m^3$ , (9);
$S$	dimensionless $s$ , (9);
$t_f$	fin thickness, m;
$t$	temperature, $^{\circ}C$ , (7);
$t_\infty$	temperature of the surroundings, $^{\circ}C$ , (7);
$T_f$	dimensionless fin thickness, $t_f/L$ ;
$T_r$	reference temperature, $^{\circ}C$ , (8);
$T$	dimensionless temperature, $t/t_\infty$ , (7);
$U$	velocity component in $x$ direction, $u/(v/L)$ , (7);
$V$	velocity component in $r$ direction, $v/(v/L)$ , (7);
$X$	axial coordinate, $x/L$ , (6);
$\beta$	volumetric coefficient of thermal expansion, $1/K$ , (8);
$\nu$	kinematic viscosity of fluid, $m^2/s$ ;
$\rho$	fluid density, $kg/m^3$ ;
$\rho_r$	reference density, $kg/m^3$ .

### Subscripts

$f$	fluid, fin;
$i$	inner cylinder;
$o$	outer cylinder;
$s$	solid;
$\infty$	surroundings.

### References

- Benjamin A S, Denny V E 1979 On the convergence of numerical solutions for two-dimensional flows in a cavity at large Re. *J. Comput. Phys.* 33: 340–358
- Bhushan R, Keyhani M, Christenen R N, Kulacki F A 1983 Correlations for convective heat transfer in vertical annular gas layers with constant heat flux on the inner wall. *ASME J. Heat Transfer* 105: 910–912
- Bozeman J D, Dalton C 1973 Numerical study of viscous flow in a cavity. *J. Comput. Phys.* 12: 348–363
- Catton I 1978 Natural convection in enclosures. *Proc. 6th Int. Heat Transfer Conf.* Toronto, Canada, 6: 13–31
- Collier J G, Boyce B E, deForge Dedman A S, Khanna R 1970 Natural convection through narrow vertical unheated annuli at high gas pressures. *Proc. 4th Int. Heat Transfer Conference*, Paris-Versailles, vol. 4, paper NC 2.6

- De Vahl Davis G 1983 Natural convection of air in a square cavity: a bench mark numerical solution. *Int. J. Numer. Methods Fluids* 3: 249–264
- De Vahl Davis G, Thomas R W 1969 Natural convection between concentric vertical cylinders, high-speed computing in fluid dynamics. *Phys. Fluids Suppl. 11* pp. 198–207
- Farouk B, Ball K S, Dixit V C 1990 Aspect and radius ratio effects on natural convection in a vertical annulus. *Proc. 9th Int. Heat Transfer Conference*, Jerusalem, Israel, vol. 2 pp. 585–590
- Ghia U, Ghia K N, Shin C T 1982 High-Re solutions for incompressible flow using the Navier–Stokes equations and a multigrid method. *J. Comput. Phys.* 48: 387–411
- Hoogendoorn C J 1986 Natural convection in enclosures. *Proc. 8th Int. Heat Transfer Conf.* San Francisco, CA, USA 1: 111–120
- Iwatsu R, Hyun J M, Kuwahara K 1990 Analyses of three-dimensional flow calculations in a driven cavity. *Fluid Dyn. Res.* 6: 91–102
- Iwatsu R, Hyun J M, Kuwahara K 1992 Numerical simulation of flows driven by a torsionally oscillating lid in a square cavity. *J. Fluids Eng.* 114: 143–151
- Iwatsu R, Ishii K, Kawamura T, Kuwahara K, Hyun J M 1989 Numerical simulation of three-dimensional flow structure in a driven-cavity. *Fluid Dyn. Res.* 5: 173–189
- Keyhani M, Kulacki F A, Christensen R N 1983 Free convection in a vertical annulus with constant heat flux on the inner wall. *ASME J. Heat Transfer* 105: 454–459
- Lin D S, Nansteel M W 1987 Natural convection in a vertical annulus containing water near the density maximum. *ASME J. Heat Transfer* 109: 899–905
- Molki M, Faghri M 1994 Conjugate natural convection heat transfer in a vertical annulus with internal circumferential fins. *Numer. Heat Transfer A25*: 457–476
- Molki M, Shahsavan M 1989 Enhancement of natural convection heat transfer in annular enclosures. *J. Eng.* 2: 45–53
- Nagendra H R, Tirunarayanan M A, Ramachandran A 1970 Free convection heat transfer in vertical annuli. *Chem. Eng. Sci.* 25: 605–610
- Ostrach S 1982 Natural convection heat transfer in cavities and cells. *Proc. 7th Int. Heat Transfer Conf.* 1: 365–379
- Ostrach S 1988 Natural convection in enclosures. *ASME J. Heat Transfer* 110: 1175–1190
- Ozawa S 1975 Numerical studies of steady flow in a two-dimensional square cavity at high Reynolds numbers. *J. Phys. Soc. Jpn.* 38: 889–895
- Patankar S V 1980 *Numerical heat transfer and fluid flow* (New York: McGraw Hill)
- Prasad V, Kulacki F A 1985 Free convective heat transfer in a liquid-filled vertical annulus. *ASME J. Heat Transfer* 107: 596–602
- Schreiber R, Keller H B 1983 Driven cavity flow by efficient numerical technique. *J. Comput. Phys.* 49: 310–333
- Sheriff N 1966 Experimental investigation of natural convection in single and multiple vertical annuli with high pressure carbon dioxide. *Proc. 3rd Int. Heat Transfer Conf.* vol. 2: 132–138
- Thomas R W, De Vahl Davis G 1970 Natural convection in annular and rectangular cavities, a numerical study. *Proc. 4th Int. Heat Transfer Conf.* vol 4, paper NC 2-4
- Tuann S Y, Olson M D 1978 Review of computing methods for recirculating flows. *J. Comput. Phys.* 29: 1–19
- Vijayan P K, Saha D, Venkat Raj V 1986 Studies on natural convection in vertical annuli heated from below. *Proc. 8th Int. Heat Transfer Conf.* San Francisco, CA, 4: 1563–1568



## Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces

C J HOOGENDOORN, C L KOSTER and J A WIERINGA

J M Burgers Centre for Fluid Mechanics, Lorentzweg 1, 2628 CJ Delft,  
The Netherlands

**Abstract.** For the combustion of natural gas in high temperature glass furnaces a computational model “Furnace” has been developed. It includes 3-D turbulent flow, flame chemistry, radiative heat transfer and the formation of soot and of the pollutant NO. Turbulent fluctuations have been taken into account, and are shown to have a large effect on thermal radiation and NO-formation. Spectral behaviour of gas radiation results in changes of heat transfer efficiency up to 5%, depending on refractory emissivity.

The model has been employed to predict NO formation for different burner geometries. In general, a decrease in mixing of gas and air results in a reduction of 1600 to 400 ppm in flue gas NO concentration. Except for some of the low mixing flames, however, they lead to a lower burnout and a very high CO level in the flue gas. A comparison with semi-technical furnace tests shows that the model can predict NO formation reasonably well. With this computational model the designer of furnaces and burners can study further possibilities for increased furnace performance and low NO emissions.

**Keywords.** Turbulent combustion; computational heat transfer; radiative heat transfer; NO<sub>x</sub> formation; glass furnaces.

### 1. Introduction

Owing to the increasing costs of energy, the higher demands on the quality of glass and the need to decrease NO<sub>x</sub> emission, the glass-industry needs more detailed information concerning the processes in glass-melting furnaces. Experimental investigations of glass-furnaces are difficult and expensive because of the high temperatures involved, typically 2000 K in the combustion chamber, so the use of models is useful. Often physical flow models have been used (Günther 1973; Gustafson 1980; Gustafson *et al* 1981, pp. 16–20). More recently also computational models have been considered (Gosman *et al* 1980; Ungan 1985). A complete description of the glass melting process requires a model for the laminar glass flow and a model for the turbulent gas flow in the combustion chamber. While the cooperating groups of Simonis *et al* (1986, 1993) and Muysenberg *et al* (1993) modelled the glass-flow,

we have concentrated on the processes in the combustion chamber for the case of natural gas-fired glass furnaces. Due to the high flame temperatures we can expect a high level of  $\text{NO}_x$  formation in such a furnace.

The use of advanced mathematical simulation models for glass furnaces have been recently reported by Gosman *et al* (1980), Carvalho *et al* (1987) Carvalho & Nogueira (1990), Ungan & Viskanta (1987). They describe the heat transfer from the flame. However, Ungan & Viskanta (1987) also give a modelling of the 3-D flow in the glass melt. A similar 3-D glass-melt model has been reported by Simonis *et al* (1986), whereas Horvath & Hilbig (1988) reported on a 2-D model coupled to a simplified flame-radiation model.

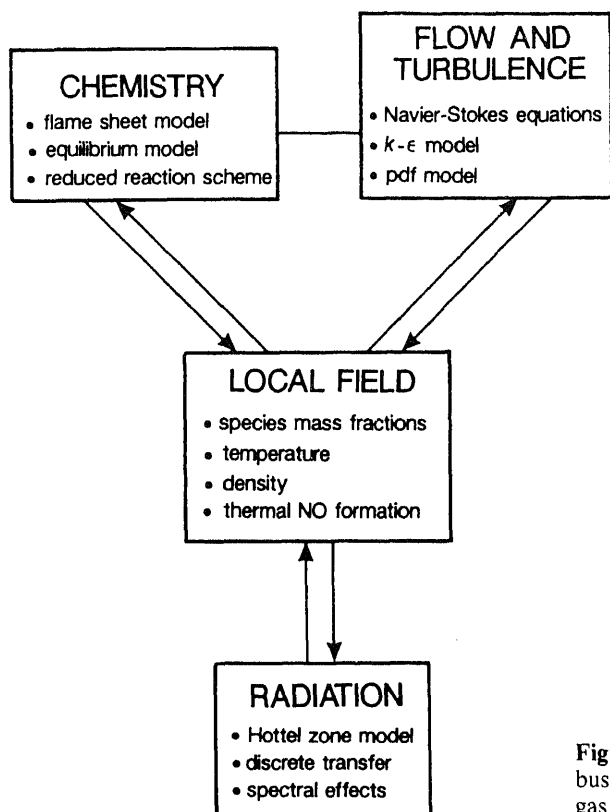
In this paper a complete model of 3-D flow and combustion, as well as of radiative heat transfer and of  $\text{NO}_x$  formation for gas-fired glass furnaces will be presented. For the 3-D turbulent flow a standard  $k-\epsilon$  model has been used. Combustion is incorporated through a simple model, using the transport equation for fuel fraction. The heat of combustion is incorporated in the energy equation. In this equation the radiative heat transfer term is also included in the source term. A problem in radiation modelling is that it essentially requires the solution of an integral equation, whereas all other equations are partial differential equations typical for transport of mass, heat and momentum. This means that the radiation term needs a separate solution procedure. Two methods will be discussed: the Hottel zone and the discrete transfer methods. Spectral radiative effects for the typical gas emission of the flue gases will be shown. In standard  $k-\epsilon$  turbulence modelling only the Reynolds averaged or Favre averaged values of transport quantities is used. However, in nonlinear source terms, as for instance radiation or chemical reactions (combustion), the fluctuating character of the turbulent fields has to be accounted for. This has been done by using an assumed shape for the probability density function of the turbulent fluctuating quantities involved. A transport equation for the variance of these fluctuations is solved to obtain quantitative results. The importance of the fluctuations will be shown for the radiative source term as well as for the combustion reactions. For the formation of NO a separate post-processor has been developed. It contains a transport equation for NO, with the thermal NO reaction as source term, via the Zel'dovich mechanism.

Computational results will be compared to experimental results obtained in a semi-technical furnace at the International Flame Research Foundation (IFRF) at Ijmuiden, The Netherlands.

## 2. The simulation model

### 2.1 Furnace code

A complete computational model for the simulation of a natural gas-fired furnace should contain all physical factors contributing to furnace performance. This means that it should contain model-flame chemistry, turbulent flow and mixing as well as heat transfer. Figure 1 shows the total scheme of the "Furnace" code of this study. The different factors interact and determine local field quantities like temperature, velocities and chemical composition of the gases. From the results the heat transfer to the glass, the emission of pollutants like  $\text{NO}_x$  and CO can be found.



**Figure 1.** Simulation of flow, combustion and heat transfer for natural gas flames.

## 2.2 The flow model

The flow model of the combustion chamber uses a numerical method to solve the continuity equation and the three-dimensional Navier–Stokes equations for a steady state case. The effect of turbulence on the flow has also to be modelled. This has been done by Launder & Spalding (1980) using a standard  $k$ – $\varepsilon$  model with wall functions. As a result a set of six coupled equations is obtained from which the three velocity components can be found, and the two variables of the turbulence model, the turbulent kinetic energy  $k$  and the dissipation of the turbulent energy  $\varepsilon$ . The general transport equation for each of these variables can be written as,

$$\text{div}(\rho \mathbf{v} \phi) = \text{div}(D_\phi \text{grad}(\phi)) + S_\phi, \quad (1)$$

where  $\rho$  = density and  $D_\phi$  = diffusivity. Here  $\phi$  stands successively for the three velocity components of the vector  $\mathbf{v}$  and for  $k$  and  $\varepsilon$ . With  $\phi = 1$  one has the continuity equation. The first term gives the convective effects ( $\text{div}(\rho \mathbf{v} \phi)$ ), the second, the diffusion effects ( $\text{div}(D_\phi \text{grad}(\phi))$ ), and the last ( $S_\phi$ ) is a source term.

### 2.3 The energy equation

The energy (enthalpy) equation has also to be solved to obtain the temperature field. This equation has the same form as (1). The source term in it comes from the enthalpy of combustion and radiative energy exchange.

### 2.4 The combustion model

When modelling combustion of natural gas a rather complex flame chemistry is involved. Very often it is sufficient for heat transfer and flow modelling to use reduced chemical reaction schemes. Also depending on reaction rate and mixing rate time scales an equilibrium model or a full kinetic model has to be used. For large furnaces at high temperatures an equilibrium or flame sheet model can be used; except for the formation of  $\text{NO}_x$ , which is reaction rate limited. In the "Furnace" code a simple one-step, a two-step and a constrained equilibrium model (Bilger & Stårner 1983) for  $\text{CH}_4$  combustion has been included as computational options.

When modelling the turbulent combustion in the natural gas/air mixture with a one-step model the following assumptions have been made:

(a) The reaction is expressed as



where  $s$  is the stoichiometric mass ratio oxidant/fuel.

(b) The reaction is infinitely fast, consequently the reaction rate is determined by the turbulent mixing of fuel and oxidant.

(c) The fuel and oxidant may coexist at the same place, but at different times (intermittency). This is due to concentration fluctuations caused by turbulence. They are accounted for by means of a probability density function. Here the probability density function is represented by a combination of two Dirac delta functions.

In the model the following five species are considered: fuel,  $\text{O}_2$ ,  $\text{N}_2$ ,  $\text{H}_2\text{O}$ ,  $\text{CO}_2$ . The composition of a fluid element can be expressed in terms of a mixture fraction  $f$ . The mixture fraction gives the concentration of the fuel relative to its initial concentration. For stoichiometric conditions in the present case  $f = 0.0612$ , whereas  $f = 0.0$  for air and  $f = 1.0$  for the fuel. Applying the above-mentioned assumptions one can find a transport equation again of the form of (1) for the mixture fraction  $f$ . In the two-step combustion model the reaction has been split at first into oxidation of  $\text{CH}_4$  to  $\text{CO}$  and  $\text{H}_2\text{O}$  and a later oxidation of  $\text{CO}$  to  $\text{CO}_2$ , when all  $\text{CH}_4$  has been converted.

Turbulent fluctuations of the mixture function  $f$  have been taken into account by a method as described by Khalil (1982), a probability density approach. In this approach one considers the probability density function ( $P(f)$ ) for the fluctuations of  $f$  for a location in the furnace. If we have a quantity  $\Phi(f)$  depending on  $f$ , we will have to find the average of  $\Phi$  from:

$$\bar{\Phi} = \int_0^1 \Phi(f) P(f) df, \quad (3)$$

which in case of a nonlinear relation  $\Phi(f)$  gives a value differing from  $\Phi(\bar{f})$ . Examples



of nonlinear behaviour are the emissive radiative power ( $T^4$ ) or the formation rate of  $\text{NO}_x$  in a furnace. Good predictions of these quantities require the use of  $P(f)$ . For the turbulent fluctuations of  $f$  we can introduce a fluctuation function  $g$  by,

$$g = (f' - \bar{f})^2, \quad (4)$$

with  $f'$  the fluctuating value of  $f$ . For this function  $g$ , Khalil (1982) also derived a transport equation like (1). For the entire set of equations see table 1. For the PDF one has to assume a shape of the function to give a good prediction. If one assumes intermittency, this function  $P(f)$  reduces to a double delta function, symmetric or clipped around  $\bar{f}$ . In the mixing zone of the flame this gives an adequate description for the combustion model. For  $\text{NO}_x$  formation also taking place downstream of the initial mixing a top hat distribution has been used in the "Furnace"-code.

**Table 1.** Transport equations used in "Furnace" code.

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j) = 0$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j \bar{u}_i) = -\frac{\partial \rho}{\partial x_i} - \frac{\partial}{\partial x_j}(\bar{\tau}_{ij}) + \bar{\rho} g_i$$

$$\tau_{ij} = -(\mu + \mu_t) \left( \frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i} \right) + \frac{2}{3} (\mu + \mu_t) \frac{\partial \bar{u}_k}{\partial x_j} \delta_{kj} + \frac{2}{3} \bar{\rho} k \delta_{ij}$$

$$k = \frac{1}{2} \overline{u'_i \cdot u'_i}, \mu_t = C_\mu \bar{\rho} k^2 / \varepsilon, \varepsilon = \frac{\mu}{\bar{\rho}} \left( \frac{\partial u'_i}{\partial x_j} \cdot \frac{\partial u'_i}{\partial x_j} \right)$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j k) = \frac{\partial}{\partial x_j} \left( \left( \mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right) - \bar{\rho} (G + \varepsilon)$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j \varepsilon) = \frac{\partial}{\partial x_j} \left( \left( \mu + \frac{\mu_t}{\sigma_\varepsilon} \right) \frac{\partial \varepsilon}{\partial x_j} \right) - \bar{\rho} \left( (\rho_{1\varepsilon} G + \rho_{2\varepsilon} \varepsilon) \frac{\varepsilon}{k} \right)$$

$$\bar{\rho} G = \left( \mu_t \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{2}{3} \bar{\rho} k \delta_{ij} \right) \frac{\partial \bar{u}_i}{\partial x_j}$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j H) = \frac{\partial}{\partial x_j} \left( \left( \frac{x}{C_p} + \frac{\mu_t}{\sigma_H} \right) \frac{\partial \bar{H}}{\partial x_j} \right) + \bar{S}_{rad}$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j \bar{Y}_i) = \frac{\partial}{\partial x_j} \left( \left( D_i + \frac{\mu_t}{\sigma_i} \right) \frac{\partial \bar{Y}_i}{\partial x_j} \right) + \bar{S}_i$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j f) = \frac{\partial}{\partial x_j} \left( \Gamma_f \frac{\partial \bar{f}}{\partial x_j} \right)$$

$$\frac{\partial}{\partial x_j}(\bar{\rho} \bar{u}_j \bar{g}) = \frac{\partial}{\partial x_j} \left( \Gamma_g \frac{\partial \bar{g}}{\partial x_j} \right) + C_{1g} \mu + \frac{\partial \bar{f}}{\partial x_j} \frac{\partial \bar{f}}{\partial x_j} - C_{2g} \bar{\rho} \varepsilon k$$

with  $C_\mu = 0.09$ ,  $C_{1\varepsilon} = 1.44$  and  $C_{2\varepsilon} = 1.92$ ;  $\sigma_k = 1.0$ ,  $\sigma_\varepsilon = 1.3$ ,  $\sigma_H = 0.9$  and  $\sigma_i = 0.9$ ;  $C_{1g} = 2.8$ ,  $C_{2g} = 2.0$ ,  $\Gamma_f = \Gamma_g = \mu_t' \sigma_f$  and  $\sigma_f = 0.9$ ;  $H$  is the enthalpy and  $Y_i$  the concentration of species  $i$ .

## 2.5 Radiative modelling

In order to evaluate the source term in the energy equation the radiative heat exchange must be modelled. One of the models used is the well-known Hottel zone method derived by Hottel & Sarofin (1967). One of the main reasons for choosing this method is the fact that in the present study the heat flux to the relatively cold and strongly absorbing glass surface is an important quantity. For this case the zone method gives very reliable results.

In the Hottel zone method one divides the combustion space into a number of equal-sized volume zones (parallelopipeds), surrounded by surface zones along the glass bath and roof. The three-dimensional space has to be divided in the model in  $n_i$  by  $n_j$  by  $n_k$  volume zones (flame length, width, height direction, respectively). The modelling consists of half the combustion chamber, using symmetry conditions along the mid-plane through the flame.

In the Hottel zone method one has to evaluate the radiation exchange factors:  $\overline{s_i s_j}$ ,  $\overline{s_i g_k}$ ,  $\overline{g_m g_k}$ . They represent the surface to surface, the surface to gas volume and volume to volume exchanges, respectively. With these factors evaluated beforehand and a calculated temperature field, one obtains for the net heat exchanges for surface zones:

$$Q_{n,j} = \sum_{i=1}^N \overline{s_i s_j} q_i^- + \sum_{k=1}^K \overline{s_i g_k} E_k - A_j q_j^-, \quad (5)$$

with  $q_i^-$  and  $q_j^-$  = leaving fluxes from areas  $A_i$  and  $A_j$ ,  $E_k$  = emitted flux from volume zone  $k$ , and for volume zones:

$$Q_{n,k} = \sum_{j=1}^N \overline{s_j g_k} q_j^- + \sum_{m=1}^K \overline{g_m g_k} E_m - 4k_k V_k E_k, \quad (6)$$

with  $k_k$  = absorption coefficient of volume  $V_k$ .

Calculation of the  $\overline{g_m g_k}$  factors is especially cumbersome as they contain a six-fold space integration. Therefore, use was made of an algorithm from Siddal (1986) for the exchange factors. Up to now the full zone method was only applied for one gray gas with a constant absorption coefficient  $k$ . For the latter a value of  $k = 0.12 \text{ m}^{-1}$  has been taken as a good average value for the whole space. This has been done on basis of a literature review for natural gas flames by Post (1988). In this reference it has also been shown that the results are not very sensitive to  $k$  in the range of 0.08 to 0.15.

Next to the Hottel method, a second radiative model has been included in the "Furnace" code as a computational option. The Discrete Transfer Method (DTM) solves the equation of radiative transfer in a different way. A detailed transcription of the DTM can be found in Lockwood & Shah (1981) and Wieringa (1992). The enclosure that is studied is divided into a number of volume and surface elements, as in the zone method. In each surface element a number of beam directions is chosen. From here each beam is traced towards the other side of the enclosure from where the beam is thought to emerge with an energy dependent on the boundary conditions. Then the beam is followed back and the intensity is determined in each volume that it passes, until the beam has reached the surface element. A major problem with the

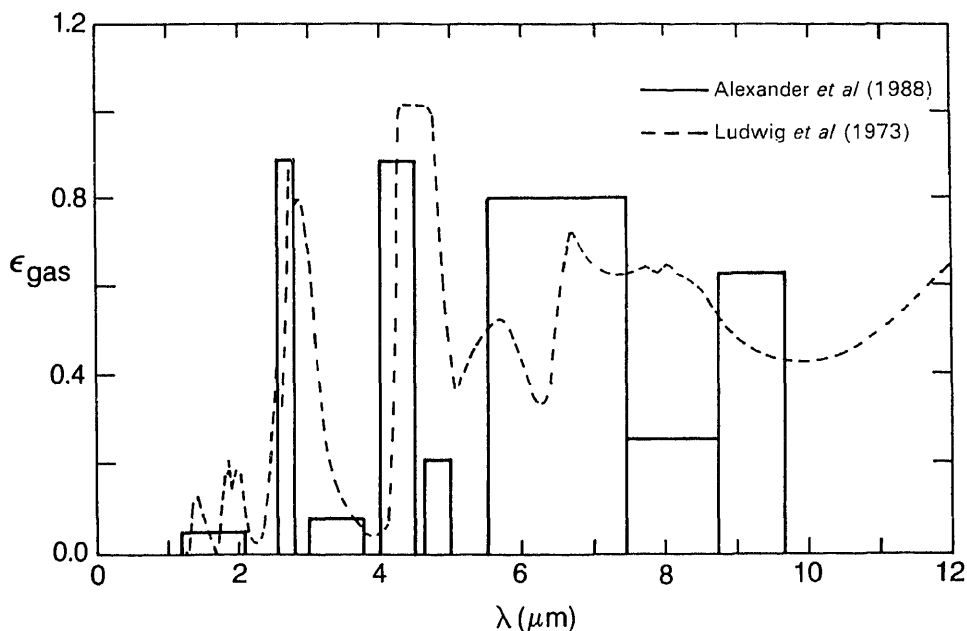


Figure 2. The two gas spectra.

DTM is the fact that a heat balance over the complete enclosure is not automatically satisfied, it requires a sufficient number of beam directions (16 or 32).

The DTM has the advantage, compared to the Hottel zone method, that the absorption coefficient can easily be made position-dependent. Hence, it is possible to take into account the effect of soot on heat transfer by radiation. The disadvantage is that the computational time per radiation iteration is higher than with the Hottel zone method.

In a few detailed studies also the spectral character of gas radiation for the combustion gases ( $\text{CO}_2$  and  $\text{H}_2\text{O}$ ) have been taken into account. Two gas spectra have been used, figure 2. The first has been taken from an earlier study by Alexander *et al* (1988) on the spectral effects in furnaces. It consists of 15 wavelength bands, of which 7 are perfectly transparent. It was computed according to data from Edwards (1960) for a somewhat smaller furnace and for gas of lower temperature. It has been used to show the influence of the combustion profile on the spectral effects, using the Hottel zone method. At 2000 K the total emissivity of the 15 band spectrum is 0.14.

The second spectrum has been computed according to the statistical narrow band model by Goody (1964). This states that the mean gas emissivity of a narrow spectral band at wave number  $\nu$  can be written as:

$$\epsilon_{\text{gas}, \nu} = 1 - \exp(-\bar{W}/\delta) \quad (7)$$

where  $\bar{W}$  is the so-called equivalent width of a spectral line, averaged over a distribution of line strengths and  $\delta$  is the mean line spacing. An exponential-tailed inverse distribution has been taken for the probability density function of line strength. Soufiani *et al* (1986) showed that this distribution gives very accurate results for spectral lines having a Lorentz profile. Ludwig *et al* (1973) give approximations for

the emissivity using various line shapes and probability density functions, and tables of experimental values for several variables in wave number intervals of  $25 \text{ cm}^{-1}$ . This way a spectrum has been obtained by using 372 intervals that are believed to reproduce the form of the spectrum accurately. It is shown in figure 2 for a mean beam length of 2.83 m and partial pressures of  $\text{CO}_2$  and  $\text{H}_2\text{O}$  of 0.08 and 0.15 atm, respectively. This spectrum yields a total emissivity of 0.23 at 2000 K.

The apparent differences between the two gas spectra can be explained from the different geometries, gas temperatures and possibly also gas compositions, and also from the fact that the 15-band spectrum has been calculated using a wide band model. It should be noticed that the maximum of the Planck curve at high temperatures in the furnace lies at about  $1.5 \mu\text{m}$ . This makes large differences, at high wavelengths, relatively unimportant.

## 2.6 Modelling of turbulence-radiation interaction

Due to turbulent fluctuations of chemical composition ( $\text{CO}_2$ ,  $\text{H}_2\text{O}$ -concentration) and of temperature, the transmissivity (absorptivity) and emissivity in a grid cell will fluctuate as well. Due to nonlinear behaviour of gas-radiation simple average values can give less accurate results. This effect has been investigated for a number of typical radiation beams for a converged Furnace-code (including grey-gas radiation) computational result using the following method.

In order to find the spectral, directional radiative flux in a point, one beam of radiation is followed in a certain direction. In the first place, we determine the volumes that are run through, and the corresponding distances. In each volume the chemical composition and temperature are determined. Then the radiative transfer equation has to be solved, which is given by:

$$L_\omega(x) = L_\omega(0)\tau_\omega(0, x) + \int_0^x e_{b,\omega}(x') \frac{\partial \tau_\omega(x', x)}{\partial x'} dx' \quad (8)$$

or in discretized form:

$$L_\omega(n) = L_\omega(0)\tau_\omega(0, n) + \sum_{m=1}^n e_{b,\omega}(m)(\tau_\omega(m, n) - \tau_\omega(m-1, n)). \quad (9)$$

Here  $\omega$  is wave number ( $1/\lambda$ ),  $L_\omega$  is spectral luminance,  $\tau_\omega$  is transmissivity and  $e_{b,\omega}$  is directional black body emissive power, given by the Planck function. Further,  $x$  is the coordinate in the given direction and  $n$  and  $m$  are volume numbers. Transmissivities from the point of observation are calculated making use of the narrow band model as described above using 372 wave number intervals of  $25 \text{ cm}^{-1}$ .

Chemical composition and temperature in a volume are found from local values of mixture fraction ( $f$ ), mixture fraction variance ( $g$ ), enthalpy ( $h$ ) and temperature ( $T$ ). Now the radiative fluxes from the turbulent flame can be modelled in three ways.

Method 1 makes use of mean values for the partial pressures and temperature. Thus, the emissive power in each volume is calculated as:

$$e_{b\omega} = e_{b,\omega}(\bar{T}). \quad (10)$$

This method clearly underestimates the emission of the gas, because of the nonlinearity of emissive power with temperature.

(2) A better way is therefore method 2, which uses pdf-averaged values for the emissive power in each volume:

$$e_{b,\omega} = \int_0^1 e_{b,\omega}(T(f))P(f)df. \quad (11)$$

This method neglects, however, the correlation between emissivity (or  $\tau$ ) and (black body) emissive power. When the instantaneous value of  $f$  is near the stoichiometric value ( $f_{st}$ ), the concentrations of  $H_2O$  and  $CO_2$  and the temperature will be high, so that at the same time emissivity and emissive power will be high. When  $f$  is far from  $f_{st}$ , the reverse is the case. In this method, however, transmissivity is calculated from mean properties, and the correlation is lost.

(3) Drawbacks of method 2 are avoided when an ensemble average is calculated of the spectral flux, using a large number of instantaneous (spatial) distributions of  $f$  along the optical path. With the double-delta pdf, it is possible to evaluate every possible  $f$ -distribution, but it appears to be more convenient when these are generated statistically. Then each instantaneous value of  $f$  in a volume is determined by a random number. The pdf prescribes the relation of  $f$  with the random number  $R$ ,

$$f = P_c^{inv}(R), \quad (12)$$

here  $P_c$  is the cumulative distribution function of  $f$ ,

$$P_c(f) = \int_0^1 P(f')df'. \quad (13)$$

Taking a large enough number of beams will give a small statistical error in the solution. In practice 250 beams were enough to obtain a standard deviation in the total flux of less than 1%. This stochastic procedure is also described by Faeth *et al* (1989, pp. 1–38).

## 2.7 $NO_x$ formation model

A thermal NO formation model has been added to the combustion model. Thermal NO formation occurs via the so-called Zel'dovich mechanism (Zel'dovich 1946):



where  $k_i$  is the Arrhenius constant of reaction  $i$ .

Reaction scheme 1 results in the following production equation:

$$\frac{d[NO]}{dt} = \frac{2[O]}{k_2[O_2] + k_3[NO]} [k_1 k_2 [N_2][O_2] - k_3 k_4 [NO]^2]. \quad (15)$$

The Arrhenius constants used for all reactions were obtained from Warnatz (1983) and Hanson & Salimian (1984) and are given in table 2. For the O-atom concentration

**Table 2.** Arrhenius constants of the reaction of the Zel'dovich mechanism.  $k_i = A \cdot T^n \exp(-E/RT)$ .

	$A$ ( $\text{m}^3/\text{kmol s}$ )	$n$ ( $-$ )	$E$ ( $\text{kJ/kmol}$ )
$k_1$	$1.8 \cdot 10^{11}$	$-$	$3.19 \cdot 10^5$
$k_2$	$6.4 \cdot 10^6$	$1$	$2.61 \cdot 10^4$
$k_3$	$4.10 \cdot 10^{10}$	$-$	$4.23 \cdot 10^3$
$k_4$	$1.37 \cdot 10^6$	$1$	$1.60 \cdot 10^5$
$K_{eq}$	$2.25 \cdot 10^4$	$-$	$4.96 \cdot 10^5$

in (15) we used the reaction,



which for equilibrium gives

$$[\text{O}] = K_{eq}^{1/2} [\text{O}_2]^{1/2}. \quad (17)$$

Since the NO-formation rate is slow compared to the combustion and the NO concentration is low relative to the concentration of other species in the flame, the NO formation is assumed not to influence combustion. This enables us to compute the NO formation in a post-processor to the combustion code, where the output of the combustion simulations can be used as input for the NO formation computations. A transport equation like (1) (see also table 1) is solved for the NO mass-fraction, in which (15) gives the source-term. A tophat PDF has been used to find the effect of turbulent fluctuations on the NO source term.

## 2.8 Soot model

The presence of soot in a furnace flame promotes radiation and hence the efficiency of heat transfer from the flame. In order to estimate this effect, the amount of soot and the distribution in the flame have to be known. Despite the highly complex nature of the chemistry of soot formation and oxidation in flames, some general principles begin to be well understood. Four more-or-less separate processes can be recognized:

*nucleation*: fuel is decomposed and the first soot nuclei appear;

*coagulation*: the soot particles collide and coalesce;

*surface growth*: the mechanism by which the bulk of the soot mass is generated;

*oxidation*: soot is oxidized mainly by  $\text{O}_2$  and  $\text{OH}$  to form gaseous products like  $\text{CO}$  and  $\text{CO}_2$ .

The amount of soot formed depends strongly on the type of fuel. Methane and ethane give relatively low soot concentrations compared to ethylene and acetylene.

The soot formation rates are kinetically limited, especially in the first part of the flame. This means that state relationships, e.g., soot mass concentrations as a function of mixture function  $f$  will depend on residence time. Accurate simulation of soot requires solving a transport equation for soot, with source terms for formation and oxidation.

A two-equation model for the evolution of soot volume fraction,  $f_v$ , and particle number density,  $n$ , is presented by Syed *et al* (1990). Two transport equations like (1) have to be solved for which the following source terms were formulated:

$$\frac{d}{dt} \left( \frac{n}{N_o} \right) = \alpha(f) - \beta(f) \left( \frac{n}{N_o} \right)^2, \quad (18)$$

$$\rho_s \frac{df_v}{dt} = \gamma(f)n + \delta(f), \quad (19)$$

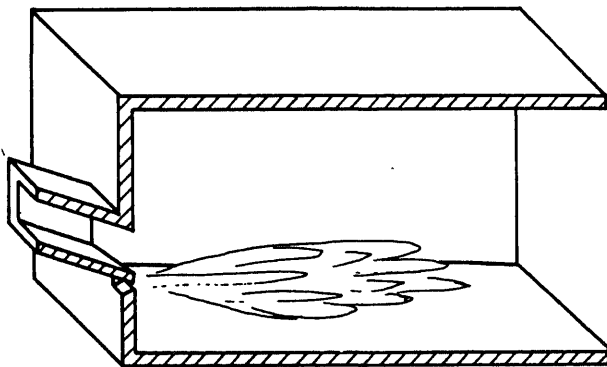
where  $\rho_s$  is the density of soot,  $N_o$  is Avogadro's number and  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  are functions of the mixture fraction which represent nucleation, coagulation, surface growth and nucleation respectively.

This model does not account for soot oxidation. However the oxidation model described by Magnussen & Hjertager (1976) was used in this study. Coupling of soot concentration to the radiation model is done by using the Planck mean of the absorption coefficient of soot.

### 3. Description of the simulated furnaces

We simulated two furnaces. One is the semi-technical size experimental furnace of the International Flame Research Foundation (IFRF) in Ijmuiden (The Netherlands). This furnace was used to simulate a glass-melting furnace, by forcing the major part of the heat loss to take place through the bottom of the furnace. In figure 3 a schematic is given of half of the furnace, with the burner port in an underport configuration. The furnace is 3.75 m long, 0.96 m high and 0.88 m wide. The heat input of the furnace is 750 kW, of which 500 kW enters via the natural gas. The temperature of the flue gas is approximately 1700 K. The gas injection velocity typically is 125 m/s, whereas the air inlet velocity is 10 m/s. The air preheat is 1373 K in the standard case.

The other furnace that we modelled is an industrial regenerative side-port fired glass-melting furnace, consisting of six burner compartments, of which three are cyclically fired. Figure 4 gives an overview of one compartment that we have modelled in our simulation studies. The dimensions of the combustion space are length 7.3 m,



**Figure 3.** View on the IFRF furnace, with underport burner system.

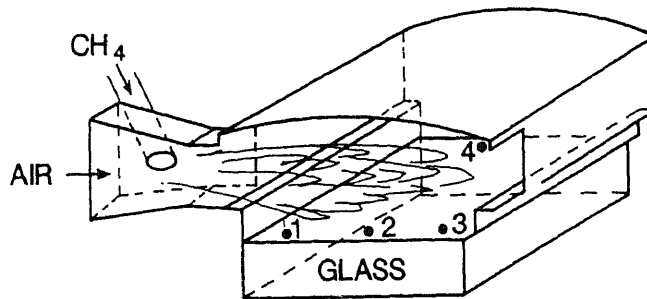


Figure 4. Regenerative side port fired glass furnace. Position of points 1 to 4, for which radiative fluxes have been calculated, are indicated.

height 2.4 m and breadth 3.3 m of one compartment. The secondary air is preheated to 1423 K, the glass bath temperature is 1773 K, air excess is 10%.

#### 4. Numerical method

##### 4.1 Solver

To solve transport equations like (1) we have used a control-volume method combined with finite differences using both the upwind and the hybrid differencing schemes. To solve the hydrodynamical problem we have used the pressure-correction method. While developing our computer program we have paid much attention to flexibility, modularity and clarity of the code. At the moment the program can use several hydrodynamical solvers (like SIMPLE, Patankar 1980, and SIMPLEST, Spalding 1980) and also several solvers for the linear equations involved. These solvers are for most variables plane-TDMA and space-TDMA for pressure correction, mixture fraction  $f$  and its fluctuations  $g$ . These two solvers are two- and three-dimensional extensions of the well-known Tri-Diagonal-Matrix-Algorithm, see Patankar (1980).

The coupled equations for flow and combustion are solved simultaneously in an iterative manner. The transport equation for  $\text{NO}_x$  however can be solved separately when a converged flow and temperature field has been obtained. The radiative heat transfer models used (Hottel and DTM) lead to integral equations for the whole furnace space based on the found temperature field. This means that in this case a transport equation cannot be used. However, in the iterative flow solutions one does not need to make a radiation calculation for every iteration. One radiation iteration suffices for every five or ten iterations.

##### 4.2 Non-rectangular geometry

As can be seen from figure 4 the roof of the combustion chamber is curved. Because the finite-differences method as used in our model is only capable of handling rectangular geometries, we had to modify our algorithm. We have chosen a porosity factor method which determines for each cell of the numerical grid the part of the cell-areas and the cell-volume available for the flow, similar to the method described by Moult *et al* (1979).



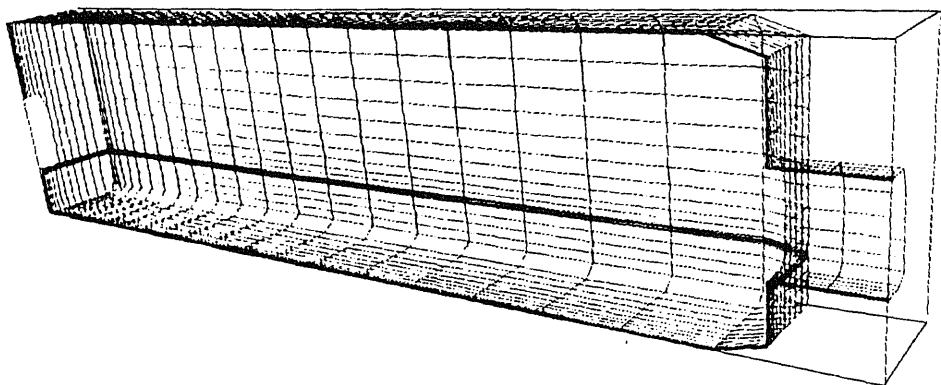


Figure 5. Grid and furnace geometry.

#### 4.3 Grid distribution

In all cases 3-D Cartesian nonlinear grids have been used. These grids are finely distributed near the inlet gas mixing zone to obtain a good prediction for the turbulent combustion field. A typical grid for the IFRF furnace with  $16 \times 24 \times 20$  points is shown in figure 5. To find grid dependency also  $32 \times 48 \times 40$  grids have been used in a few cases, this gave only slightly ( $\leq 3\%$ ) differing results.

For the radiation calculations the volume zones of the Hottel or DTM model do not have to coincide with the finite difference grid. In general we used larger gridsizes for the radiative heat transfer. The radiation source term in the energy transport equations for a finer flow field grid cell uses the contribution from the radiative power and absorption obtained from the coarser grid. We attributed to the distribution of these a weighting factor proportional to the grid volume for the absorption of radiation and to  $T_i^4 = (T_i \text{ gridpoint temperature})$  and volume for the emission of radiation. Typically one completely converged solution on the standard grid takes about 6 h of CPU-time on an HP735 computer.

#### 4.4 Validation of the model

Next to the grid refinement study as mentioned under § 4.3, the model has been validated by experiments. For the 3-D flow and mixing of natural gas and air an isothermal experiment has been carried out on a cold (air) furnace model. The flow field and the mixing of a jet of air with an He-tracer and the primary air could well be predicted (Post & Hoogendoorn 1987; Post 1988). This validated the  $k-\epsilon$  turbulence modelling for this particular flow configuration of the furnaces studied. The full experimental validation of the complete model will be discussed by comparing with measurements of a semi-technical scale test furnace.

### 5. Simulations

#### 5.1 Flow and combustion in test-furnace

Some typical calculated fields in the vertical mid-plane through the burner for the IFRF furnace are shown in figures 6 and 7. In figure 6 the velocity vector field has

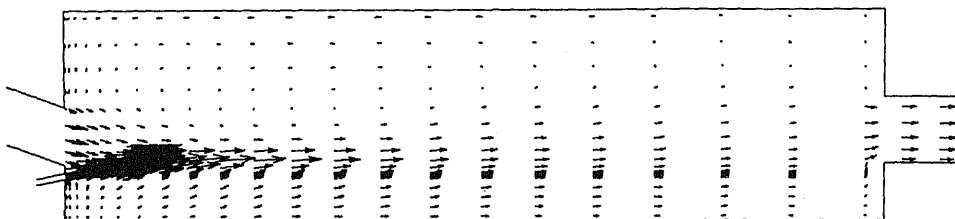


Figure 6. Velocity vectors in the symmetry plane with 12° underport firing.

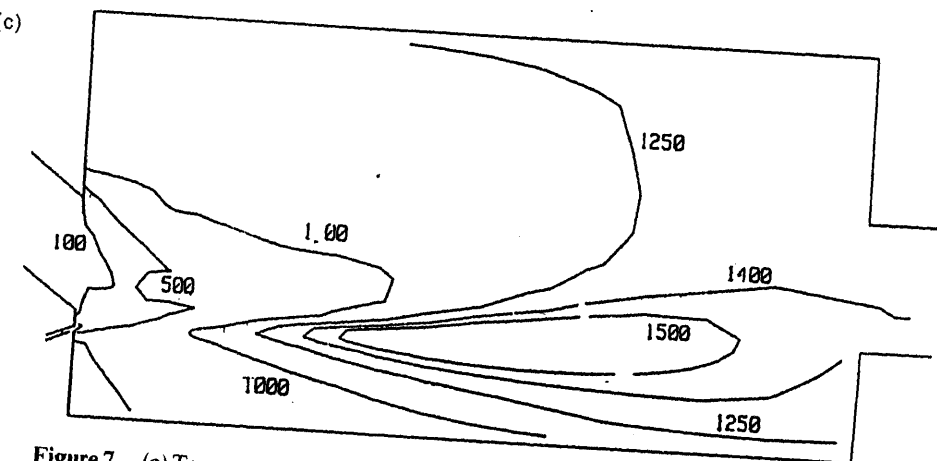
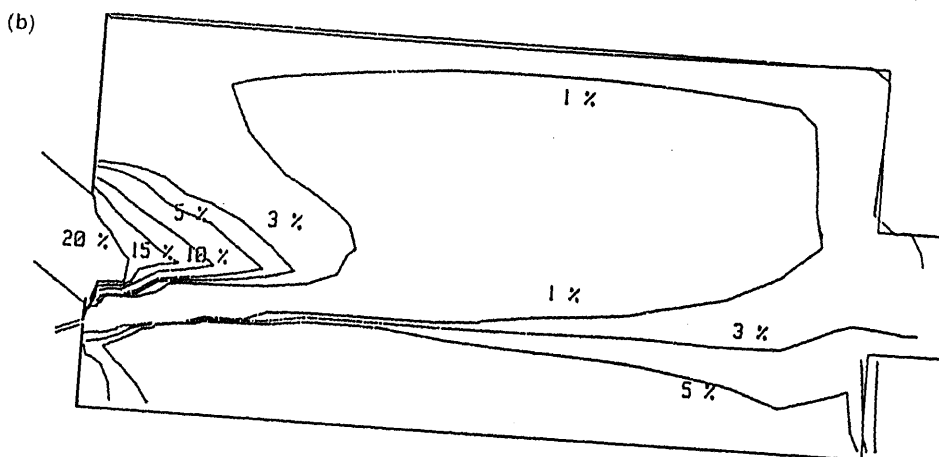
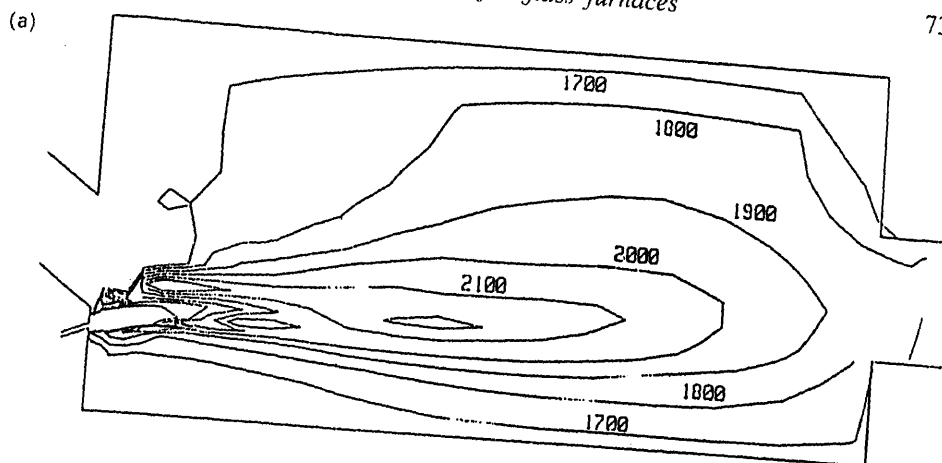
been shown. It shows a recirculation zone with low velocities above the main flame jet. Figures 7a, b and c show respectively the isotherms, lines of constant oxygen concentration and constant NO concentrations. The distribution of the radiative heat flux to the load has also been computed and figure 8 shows such a distribution for the 12° underport firing case.

The tests at the IFRF by Van de Kamp *et al* (1991) resulted in two baseline flames that were thoroughly tested during two trial periods. These two flames are both underport fired, at two different gas injection angles (12° and 20°). Plots of the temperature in the symmetry plane for both flames as simulated are shown in figure 9. It is clear from this figure that the mixing pattern is completely different for the two baseline flames. The 20° firing mode results in a much higher mixing rate and thus in faster combustion. As a result the maximum temperature is higher than in the 12° firing model, see also table 3. The predicted NO<sub>x</sub> as simulated by FURNACE is also higher: 1074 ppm at 12° against 1470 ppm at 20° (both normalized at 0% O<sub>2</sub>). The measured fluegas NO<sub>x</sub> concentrations are 1070 ppm and 1170 ppm (0% O<sub>2</sub>), at 12° and 20° natural gas injection respectively. Further, the two-step combustion model, extended for combustion to include CO formation and for NO<sub>x</sub> formation with a double delta-to-phat pdf, has been used.

A comparison can be made with the test at IFRF. In figure 8 the measured heat flux distribution is compared to the simulated one. The agreement is good only near the very end of the furnace, where the measurements show a lower flux, probably due to high radiative heat losses through the outlet and back-wall in the test. Flue gas temperatures (table 3) for two tested firing modes (12° and 20° gas injection angles) differ by 50 and 20 K for the two cases. In view of the measuring accuracy (50 K) of this temperature, good agreement is obtained, meaning that energy efficiency can be well predicted with the "Furnace" code. Taking the accuracy of the NO<sub>x</sub> measurements into account (± 150 ppm), the two-step combustion model predicts the NO<sub>x</sub> concentrations reasonably well (see also table 3).

To achieve the temperatures normal in glass-melting furnaces, the combustion air is preheated. The preheat temperature normally is approximately 1373 K. The effect of the preheat temperature on the NO<sub>x</sub> formation is illustrated in figure 10. At 1573 K air preheat the mean temperature in the furnace is 1800 K, while at 1073 K preheat the mean temperature is 1670 K.

It can be seen from figure 10 that reducing the flame temperature by reducing the air preheat seems a feasible way of reducing NO<sub>x</sub> emissions. This is confirmed by the measurements with which the simulations again agree very well. However, the furnace efficiency also decreases significantly when the air preheat temperature is



**Figure 7.** (a) Temperature, (b) oxygen and (c) NO concentration fields in symmetry plane with 12° underport firing. Temperature in K, O<sub>2</sub> concentrations in % mass fractions, O<sub>2</sub> and NO in ppm. (Horizontal dimension has been scaled down twice as much as vertical dimension.)

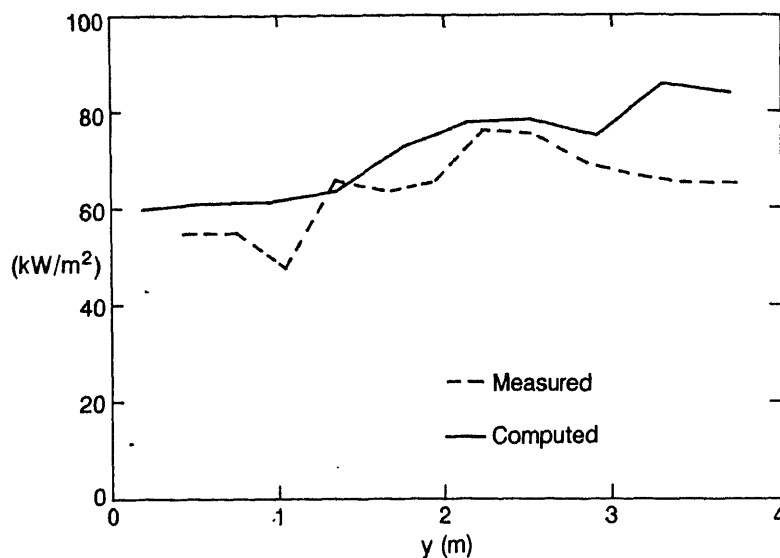


Figure 8. Computed and measured heat flux distributions in axial direction.

lowered. This results in a decrease in the heat transfer to the load of 30%, between 1573 K and 1073 K combustion air preheat. At the low preheat temperature the process may no longer be economically feasible. It should also be noted that the process requirements should always be met. Glass melting for high quality glass requires refining and is essentially achieved at high temperatures.

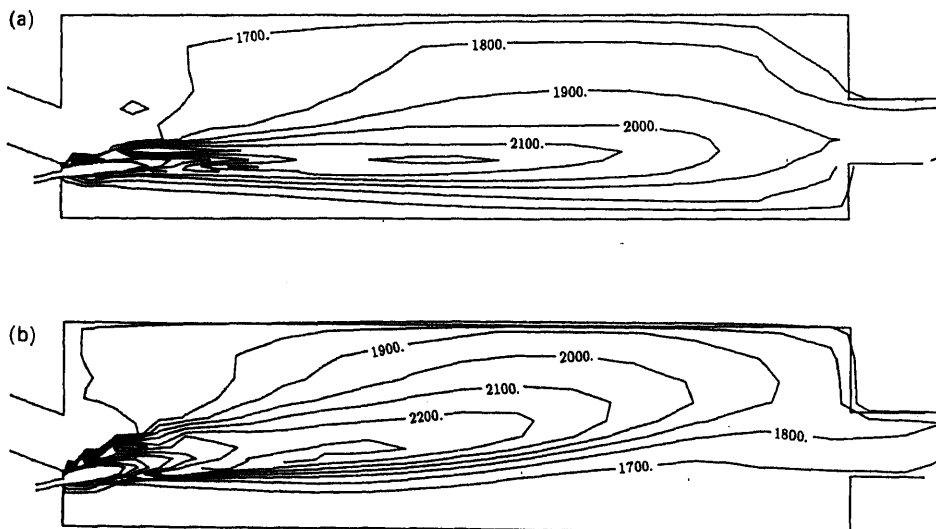


Figure 9. Temperature in the symmetry plane for the baseline flames: (a) 12° firing mode, (b) 20° firing mode.

**Table 3.** Some characteristics of the baseline 12° and 20° underport flames. Measured data from IFRF. Simulations from FURNACE and complete model.

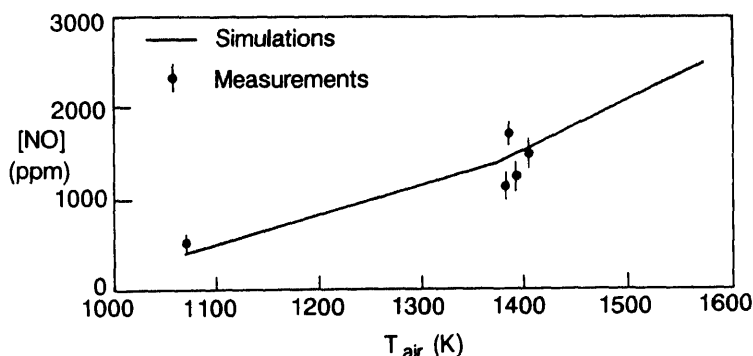
	Simulated		2-Step combination model		Measured (IFRF)	
	12°	20°	12°	20°	12°	20°
$T_{\max}$ (K)	2395	2450	2330	2390	—	—
$T_{fluc}$ (K)	1765	1710	1767	1705	1718	1685
$q''_w$ (kW/m <sup>2</sup> )	72.6	73.3	70.7	74.2	68.8	70.1
$T_{fluc}$ NO <sub>x</sub> (ppm)	1074	1470	930	1020	1070	1170

ss melting furnaces are normally operated with an excess air level between 5%–10% (airfactor  $\lambda = 1.05$ –1.30). The glass refinement requires that the molten glass be in an oxidizing environment, but the furnace efficiency requires an airfactor as low as possible, while still maintaining full burnout of the natural gas. The effect of airfactor on the NO<sub>x</sub> emission is shown in figure 11, in this case also there is agreement with measurements.

Higher airfactors result in somewhat higher NO<sub>x</sub> emissions due to the increasing oxygen concentration. Of course at high airfactors the temperature in the furnace is reduced, due to the mixing of large amounts of relatively cold combustion air. The simulations again predict the measured NO<sub>x</sub> emissions quite well. Keeping the airfactor as low as possible is a way both to reduce the NO<sub>x</sub> emissions and to increase furnace efficiency. In general we found that a less intense mixing of natural gas and air results in longer flames with less NO formation.

### Results for industrial furnace

From the full simulation model of the industrial combustion chamber details of the combustion and heat transfer can be obtained (see Post 1987, pp. 884–895, and Post & Hoogendoorn 1987). Whether the flame is short or long depends



**Figure 10.** Flue gas NO at several combustion air preheat temperatures.

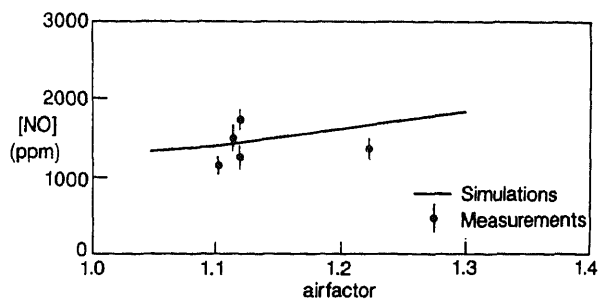


Figure 11. Flue gas NO at several airfactors.

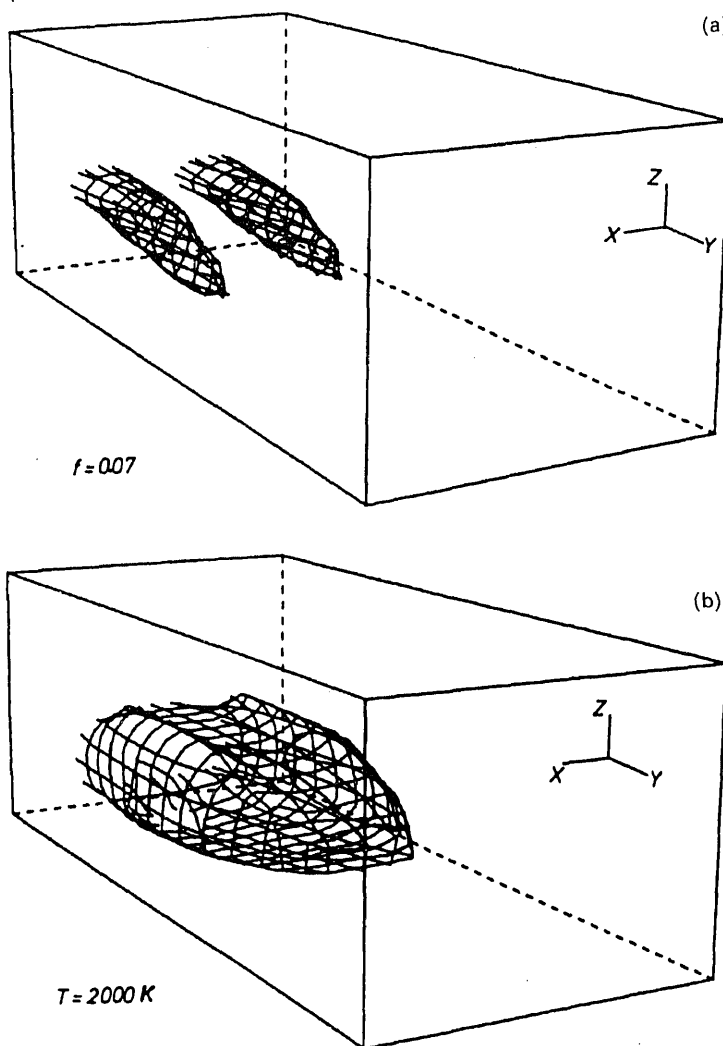


Figure 12. Flame pictures in the combustion chamber. (a) Surface of constant mixture fraction  $f = 0.07$ , showing presence of fuel ( $f = 0.0612$  stoichiometric); (b) surface at constant temperature ( $T = 2000\text{ K}$ ).

strongly on the inlet conditions. It was found that the initial turbulent length scale is particularly important here. An intense mixing with a large turbulent length scale leads to high  $\mu_t$  values and consequently to short flames. In this way the flame configuration can be influenced strongly.

Some results for a long flame (small turbulent length scale at inlet port) will now be given. A picture of a long flame can be seen in figure 12a, here the surface of a constant mixture fraction  $f = 0.07$ , corresponding to an air factor of 0.87, is shown and in figure 12b the 2000 K surface is shown. The prescribed flow profiles at the combustion chamber are for the case of a precombustion of 44% in the burner. From the computations a 97% burnout of the fuel at the outlet is found for the long flame, complete burnout will occur in the passage to the regenerator. For the short flame a 100% burnout has been found. For this hot flame an average heat flow for half of the compartment is found: to glass bath 1.51 MW (99% of this by radiation and only 1% by means of convection), wall losses 0.10 MW and flue gas heat loss 2.20 MW. The efficiency of the furnace, related to total fuel input, is 42% in this case. For a long, relatively cool flame with the same heat input the heat flow to the glass bath is 1.25 MW and the flue gas heat flow is 2.46 MW, where the flue gas still contains unburnt fuel. Combustion chamber efficiency is now 35%.

A short or a long flame also has a large effect on NO-formation. For a short flame the NO-combustion in the outlet is 3481 ppm, whereas for the long flame a value of 4007 has been found. In general these NO-values for the industrial furnace are higher than those for the IFRF-test furnace. The temperatures in the industrial furnace are higher (flue gas temperatures of 1900 K versus 1750 K). For this furnace we could clearly show the large effect of turbulent fluctuations on NO-formation. For the short flame the value of 3481 ppm drops to 1974 ppm (43% lower) if one neglects the turbulent fluctuations (tophat pdf) on the NO-source term.

### 5.3 Effects of spectral radiation

The main mode of heat transfer to the glass in a high temperature furnace is radiative heat transfer which in general is over 95% of the total heat flux. Previously Post & Hoogendoorn (1987) have shown that for prediction of fuel efficiency a simple single zone or axial flow model is sufficient. To find the effects of the spectral behaviour of gas radiation we included this in a simple model (see Hoogendoorn *et al* 1990). In this model we have the gas multiband radiation spectrum for  $\text{CO}_2$  and  $\text{H}_2\text{O}$  at temperatures and partial pressures as occurring in the furnace. For the refractory material a grey surface has been assumed. Most of these materials have a low emissivity at high temperatures. For some specific samples an emissivity of 0.4 at 2000 K has been measured by us. For the glass a grey emissivity of 0.8 has been taken.

Some results for the industrial furnace are given in figure 13. Here the effect of refractory emissivity  $\epsilon_r$  has been given. Increasing  $\epsilon_r$  results in a higher total heat flux to the glass  $Q_{gl}$ . This is due to the fact that a higher roof emissivity results in a strong increase of the roof radiation because the roof refractory transforms the banded gas radiation to a continuous emitted radiation. However, at low  $\epsilon_r$  the roof reflects the banded radiation for a large part  $(1 - \epsilon_r)$  in the same bands as the gas radiation. In these bands the gases have low transmission and will absorb the reflected radiation to a large extent before it reaches the glass. Figure 13 shows that the direct radiation decreases with higher  $\epsilon_r$ , due to the somewhat lower gas temperatures, however, the

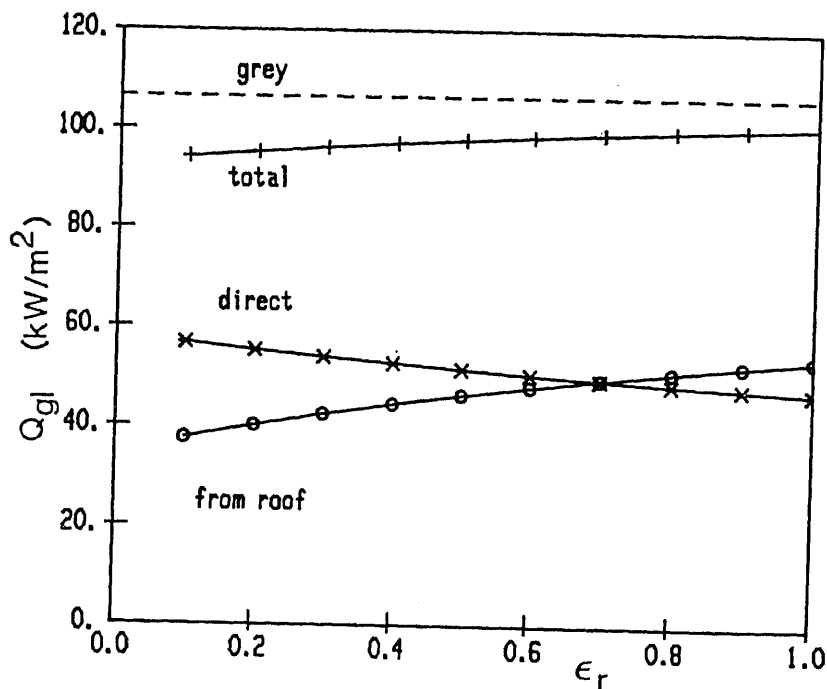


Figure 13. Heat flux to glass as function of refractory emissivity. Full lines are from spectral calculations, dotted for a grey gas.

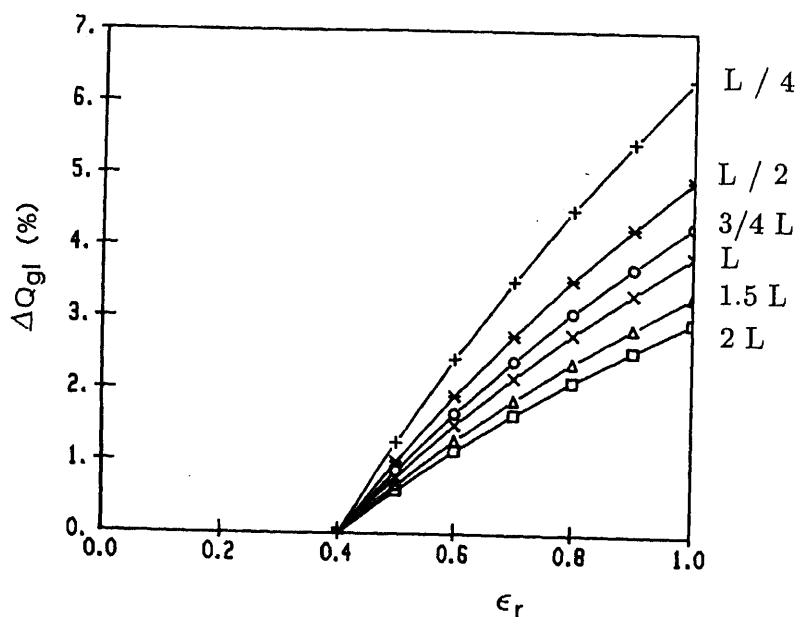


Figure 14. Relative enhancement of heat flux to glass as function of refractory emissivity, relative to  $\epsilon_r = 0.4$ . The mean length  $L = 2.8$  m.



net flux (total) still increases. The increase of the roof emissivity  $\epsilon_r$  can be obtained by structuring the refractory with cavities and protrusions to make it a "rough" surface or to apply a special coating, see Alexander *et al* (1988) and Fisher (1986). The relative importance of the spectral effects depend on the size of the furnace as expressed by the mean beam length  $L$ . Figure 14 gives results of the relative increase of glass load at constant heat input for different values of  $L$ . Small furnaces show increases of heat fluxes up to 5%, large furnaces show smaller (3%) benefits of increasing  $\epsilon_r$  from 0.4 to 1.

#### 5.4 Turbulence-radiation interaction

In order to estimate effects of turbulence on radiative fluxes in a glass furnace, a number of computations have been done for the industrial furnace. At four locations at the glass melt surface, see figure 4, the calculated spectral incident fluxes will be shown. Three of the points (1–3) lie at the symmetry plane in the centre of the flame. Point 4 is localized at the assumed symmetry plane between two flames. Vertical heights at these points are 1.9, 2.4, 2.0 and 2.4 m for points 1 to 4 respectively. In the vertical direction, the combustion chamber has been divided into 10 volumes. Because of symmetry properties, it is only necessary to consider half a burner compartment of the sideport-fired furnace. In this segment the thermal input of natural gas and preheated secondary air (1423 K) is 3.8 MW. The precombustion of the flame is 44%.

For each point,  $\bar{T}$ ,  $f$ ,  $g$  and mole fractions  $X_{H_2O}$  and  $X_{CO_2}$  are shown in figure 15. It is clear that for point 1, near the inflow opening, the path is the least uniform and the mixture fraction fluctuations  $g$  are the largest. It may therefore be expected that differences between the three computation methods will be the largest in point 1.

Table 4 gives the total (normal) fluxes from the flame for the four points and for methods 1 to 3, (10)–(13). Results of method 3 (the stochastic method) have a statistical uncertainty, for which the standard deviation is also given. Several things may be noticed from table 4. In the first place, the total flux shows a maximum somewhere halfway in the flame. Towards the end of the flame and away from the flame axis, the flux decreases.

Secondly, the differences between the fluxes calculated according to the three methods are the largest near the inflow. In point 1, the stochastic method predicts

**Table 4.** Mean total normal radiative fluxes from the combustion gases at points 1 to 4 in the glass furnace, calculated according to methods 1 to 3. For method 3, the standard deviation  $s$  in the mean flux is also given.

Point	$q(\text{kW m}^2 \text{sr}^{-1})$			$s(q_3)$ ( $\text{kW m}^2 \text{sr}^{-1}$ )
	Method 1	Method 2	Method 3	
1	51.49	53.02	65.38	0.34
2	67.05	67.48	70.33	0.26
3	54.89	55.12	56.16	0.19
4	55.39	55.77	57.52	0.18

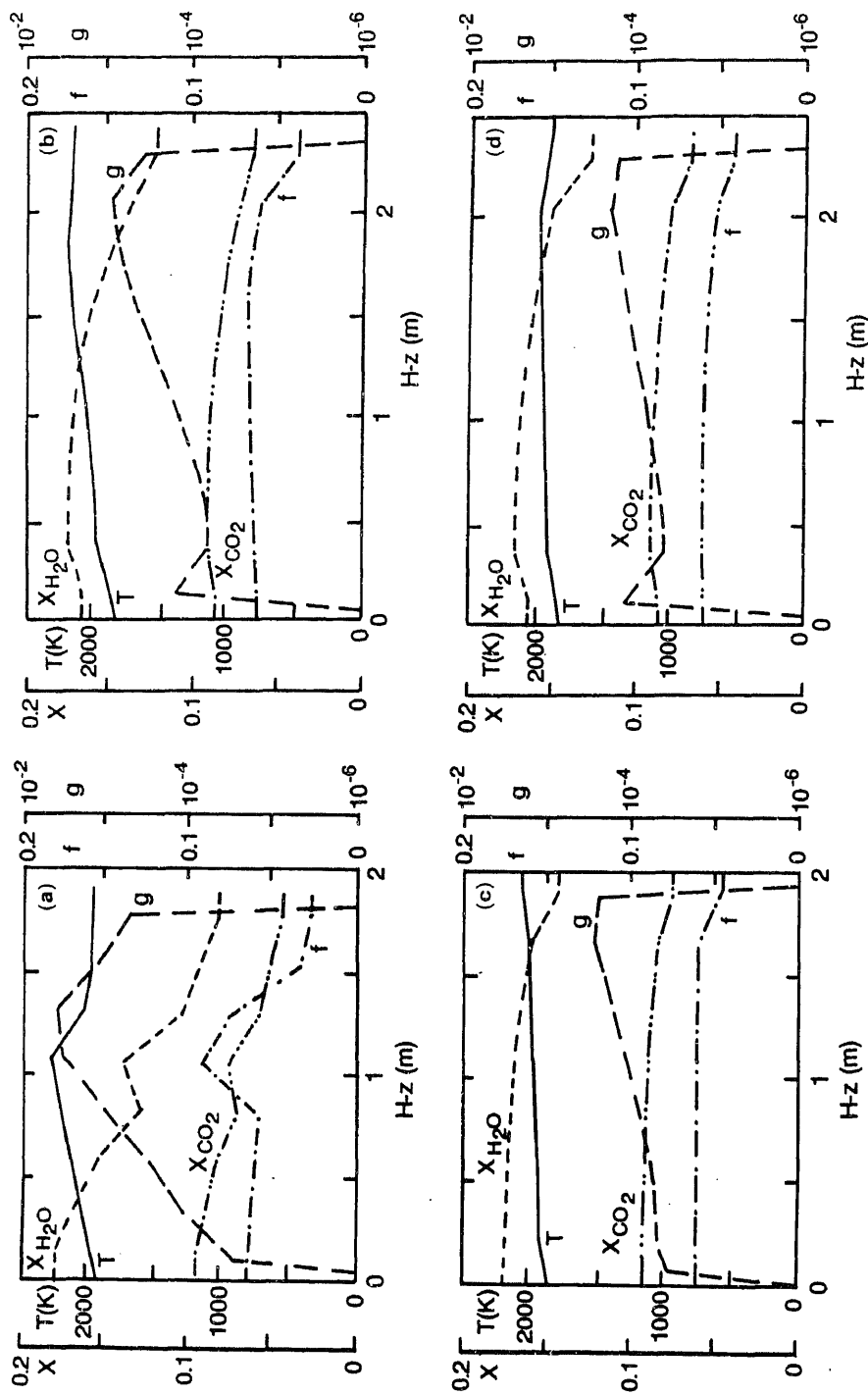


Figure 15. Mean temperature ( $T$ ), mixture fraction ( $f$ ), mixture fraction variation ( $g$ ), and mole fractions ( $X$ ) of  $H_2O$  and  $CO_2$  vs. distance from the refractory in normal direction above points 1 to 4, (a) to (d).

**Table 5.** Mean total directional radiative fluxes at point 1 in furnace, for different directions according to methods 1 to 3 as well as one case with a black background at temperature  $T_b$ .

$\theta(^{\circ})$	$T_b(K)$	$q(kW m^2 sr^{-1})$			$s(q_3)$ ( $kW m^2 sr^{-1}$ )
		Method 1	Method 2	Method 3	
+30	0	63.43	64.84	75.47	0.60
+60	0	82.05	83.82	96.30	0.60
-30	0	41.12	42.95	58.25	0.37
-60	0	18.78	20.40	35.20	0.32
0	1850	210.1	211.7	222.9	0.48

a 27% higher flux than method 1, whereas method 2 only gives a 3.0% higher flux. However, further away from the inflow opening, the effect of turbulence quickly becomes smaller. For point 2,  $q_3$  is only 4.9% higher than  $q_1$ , and in point 3 this difference is only 2.3%. Even method 3 predicts flux increases by turbulence that are lower than are given by Faeth *et al* (1989). This may be a consequence of the fact that they studied vertical diffusion flames where buoyancy can induce strong fluctuations.

Table 5 shows total fluxes in point 1, calculated for different angles with the normal. Positive values of the angle indicate that observation is in the direction of the outflow opening in the symmetry plane of the flame (figure 4). A negative value indicates a beam towards the inflow port. With increasing positive angle, fluxes are found to increase, partly because the path length is increased. Moreover, the influence of turbulence is found to decrease. The latter probably is the result of decreasing values of  $g$  towards the end of the flame. The opposite is valid for negative angles.

Also in table 5 fluxes are shown in normal direction in point 1, from the flame with a background flux. The background has been assumed to be a black surface at

**Table 6.** Influence of soot on model predictions. Soot model of Syed *et al* (1990) has been used. Different values for the constants in the soot model were used for the two different flames to obtain agreement with measurements.

	$T_{max}$ (K)	$T_{flue}$ (K)	$NO_{flue}(0\%)$ (ppm)	$Q_{load}$ ( $kW/m^2$ )	$f_{v,max}$ ( $\times 10^7$ )
<i>Underport firing</i>					
no soot	2451	1710	1408	72.25	—
with soot	2451	1701	1314	73.19	1.15
<i>Overport firing</i>					
no soot	2331	1759	1014	71.97	—
with soot	2324	1699	623	76.42	5.39

1850 K. This temperature is approximately the refractory temperature in the furnace. The refractory has been taken as black to avoid severe complication of the program when the calculation of reflected radiation is included. It turns out that with a black background radiation, the effect of turbulence is unchanged in the absolute sense. In relation with the total flux, the relative effect is reduced.

### 5.5 Effect of soot formation

Based on spectral radiation measurements carried out at the IFRF, Wieringa (1992) has estimated the amount of soot present in the IFRF furnace for two different flames. For the underport flame the maximum  $f_v$  is of the order  $1.3 \times 10^{-7}$ ; for the overport flame  $f_{v, \max}$  is of the order  $5.0 \times 10^{-7}$ .

The constants of the soot model were adjusted for each of the two flames such that the simulated values of  $f_v$  correspond to the experimental values. Table 6 shows the influence of taking soot into account in the simulations. For the underport flame, soot loadings are too low to observe any important effect. The mean temperature decreases slightly and the NO emission level reduces only 7%. However, for the overport flame the temperatures decrease significantly because the flame radiation is stronger. This has the effect that the NO emission level is reduced almost 40%. Figures 16a and b show the [NO] distribution for the overport flame with and without accounting for the effect of soot. Also the flux to the load has increased. Measured NO emission levels are 1214 ppm for the underport flame and 342 ppm for the overport flame.

### 5.6 Reduction of NO-formation

In the IFRF semi-technical furnace many tests have been done on burner modification to the effect on NO-formation and combustion. Next to underport firing, side-port and overport firing have also been done.

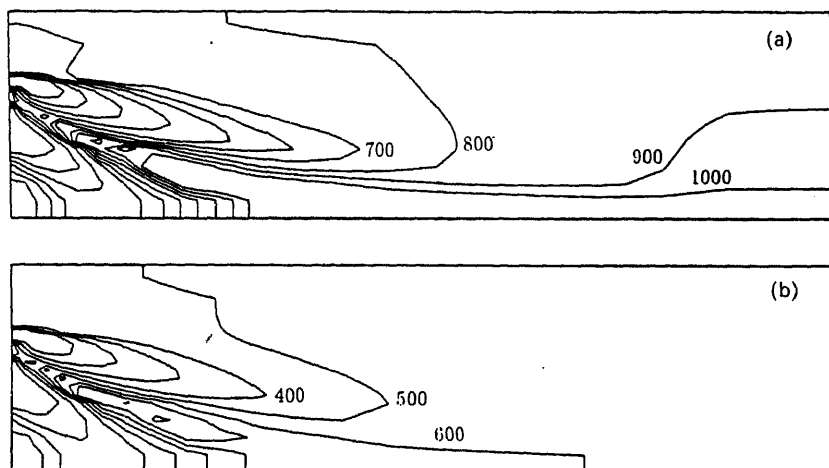
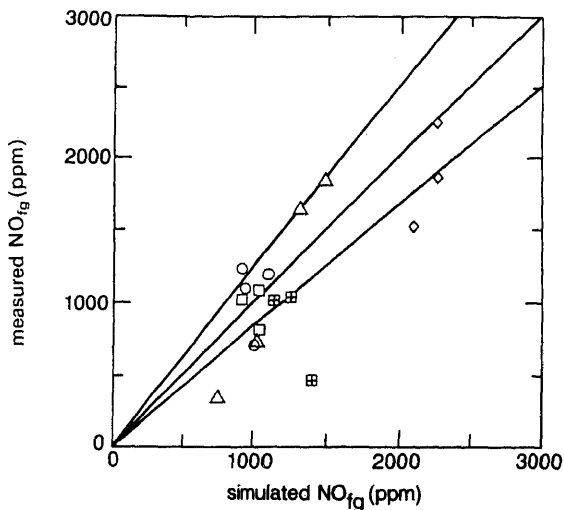


Figure 16. NO concentration (ppm) in overport flame without (a) and with (b) soot. Interval between contourlines – 100 ppm.



**Figure 17.** Comparison of the simulated and the measured NO emissions of the IFRF tests. (Circles correspond to underport flames, triangles to overport flames, squares to underport/overport flames, diamonds to sideport flames and crossed squares to parallel sideport flames. Also shown are the  $\pm 20\%$  deviation lines.)

In general, these tests resulted in lower  $\text{NO}_x$  concentrations in the flue gases, than for underport firing (see Koster 1992). The injection angle of the natural gas jet into the air flow is important. A large angle enhances mixing of gas and air leading to shorter flames but higher  $\text{NO}_x$  values. In general, a less intense mixing results in longer flames with less  $\text{NO}$ -formation. A reduction from 1600 to 400 ppm can be obtained by burner modifications. All tested cases have also been simulated (Koster 1992). Figure 17 shows that within 20% for most tests a good agreement is reached between measurements and simulations, except for a few cases of low  $\text{NO}_x$  concentrations in the tests. These were the slowly mixing, parallel-fired burner modes. In the simulations we obtained higher  $\text{NO}_x$  values, probably due to enhanced numerical diffusion in the initial part of the flame for the simulations. However, for practical applications these slowly mixing flames are unacceptable, as they give rise to a low burn-out and very high CO concentration.

Further possibilities to reduce  $\text{NO}$ -emissions like air- or gas-staging, multiple gas injectors and increased recirculations of the gases can easily be investigated with the simulation model.

## 6. Conclusions

A complete computational model for turbulent flow, combustion and heat transfer in high temperature furnaces has been developed. It includes radiative heat transfer from combustion gases as well as turbulent flow and flame chemistry. With this model furnace performance can be predicted.

Comparison of computed results with measurements in a semi-technical furnace shows good agreement for heat-flux distribution to load, furnace efficiency and concentration of the pollutant  $\text{NO}$  in flue gases. With the model many burner configurations could be computed. For the  $\text{NO}$ -concentration it could be shown that a low-mixing flame can result in a low  $\text{NO}$ -concentration.

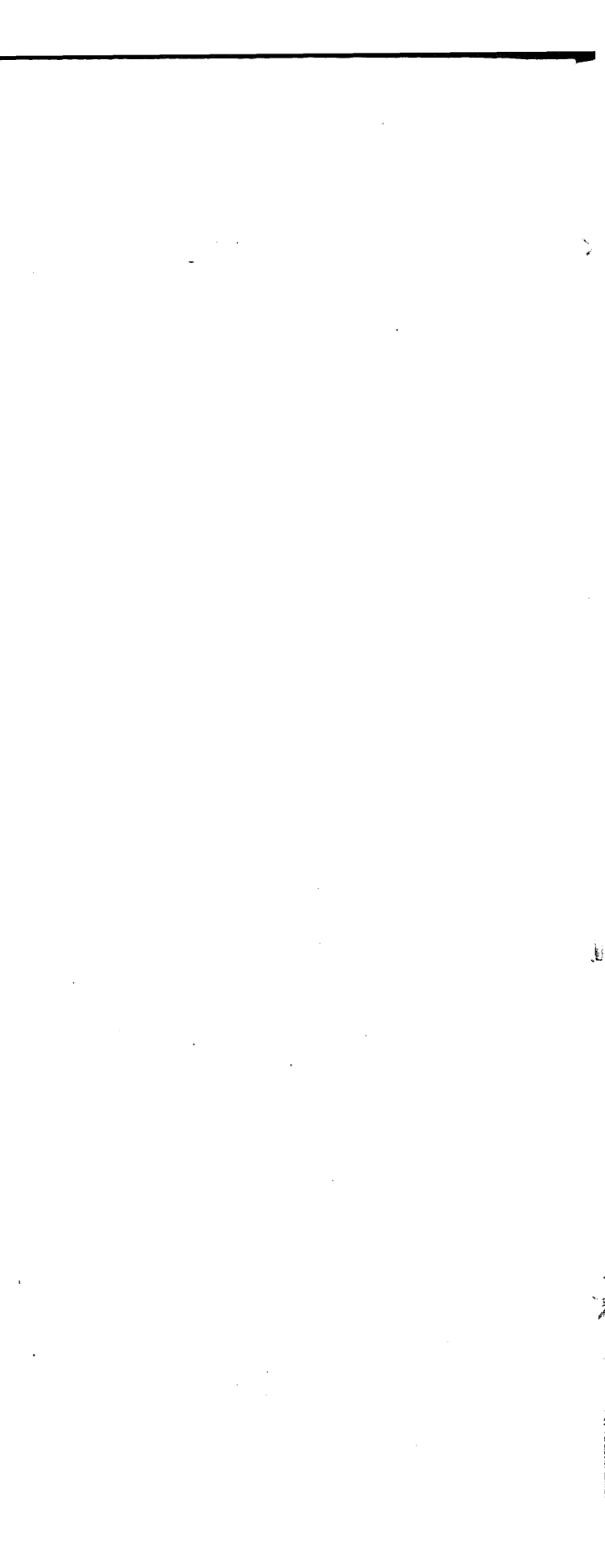
Thermal radiation is the main mode (95%) of heat transfer in these furnaces. Spectral effects of gas radiation are important. Due to this, increasing the refractory emissivity from the usual value of 0.4 to 0.95 can result in 5% increase in furnace efficiency.

Turbulent fluctuations are shown to have a large influence on the radiative heat transfer in the flame zone of the combustion chamber. These fluctuations are also important in the NO-formation. Neglecting them leads to a 43% lower value for the emission of NO in the flue gases. The developed model ("Furnace"-code) can easily be used for the design of new and improved burners or furnaces.

## References

- Alexander I, Gray W A, Hampartsoumian E, Taylor J M 1988 Surface emissivities of furnace linings and their effect on heat transfer in an enclosure. *Proc. 1st European Conf. on Industrial Furnaces and Boilers*, Lisbon
- Bilger R W, Starner S H 1983 A simple method for carbon monoxide in laminar and turbulent hydrocarbon diffusion flames. *Combustion Flame* 51: 155-176
- Carvalho M G, Durao D F G, Pereira P C F 1987 Prediction of the flow, reaction and heat transfer in an oxy-fuel glass-furnace. *Eng. Comput.* 4: 23-34
- Carvalho M G, Nogueira M 1990 Mathematical modelling of heat transfer in an industrial glass furnace. Heat transfer in radiating and combusting systems. *Proc. of Eurotherm Seminar 17* (ed.) M G Carvalho, F Lockwood, J Taine (Lisbon: Springer) pp. 374-392
- Edwards D K 1960 Absorption of infrared bands by carbon dioxide gas at elevated pressures and temperatures. *J. Opt. Soc. Am.* 50: 617-626
- Faeth G M, Gore J P, Chuech S G, Jeng S -M 1989 Radiation from turbulent flames. *Annual Review Numerical Fluid Mechanics and Heat Transfer* (eds) C L Tien, T C Chawla (New York: Hemisphere) ch. 1 p. 1-38
- Fisher G 1986 Ceramic coatings enhance performance engineering. *Am. Ceram. Soc. Bull.* 65: 284-287
- Goody R M 1964 *Atmospheric radiation I, theoretical basis* (Oxford: Clarendon)
- Gosman A D, Lockwood F C, Megahed I E A, Shah N G 1980 The prediction of flow, reaction and heat transfer in the combustion chamber of a glass furnace. *AIAA 18th Aerospace Sciences Meeting*, AIAA-80-0016
- Günther R 1973 Modelluntersuchungen über Erdgasbrenner für Wannenöfen. *Glastech. Ber.* 5: 92-98
- Gustafson E R 1980 Design study of glass manufacturing furnace using laboratory model. Sibley School of Mech. & Aerosp. Eng, Cornell University
- Gustafson E R, Torrance K E, Hayes G F, Simmons D A 1981 Laboratory model study, burner locations. *The Glassindustry* 62(5): 16-20
- Hanson R K, Salimian S 1984 Survey of rate constants in the N/H/O system. *Combustion chemistry* (ed.) W C Gardiner (Berlin: Springer Verlag) chap. 6
- Hoogendoorn C J, Post L, Wierenga J A 1990 Modeling of combustion and heat transfer in glass furnaces. *Glastech. Ber.* 65: 7-12
- Horvath Z, Hilbig G 1988 Mathematical model for fuel-heated glass melting tanks. *Glastech. Ber.* 61: 277-282
- Hottel H C, Sarofin A F 1967 *Radiative transfer* (New York: McGraw-Hill)
- Khalil E E 1982 Modelling of furnaces and combustions. *Energy and Engineering Science Series* (Kent: Abacus)
- Koster C L 1992 *Modelling of NO<sub>x</sub> formation in a high temperature gas-fired furnace*, Ph D thesis, University of Technology, Delft
- Launder B E, Spalding D B 1980 *Mathematical models of turbulence* (London: Academic Press)
- Lockwood F C, Shah N G 1981 A new radiation solution method for incorporation in general combustion prediction procedures. *Proc. 18th Symp. (Int.) on Combustion* (Philadelphia: The Combustion Institute) pp. 1405-1414

- Ludwig C B, Malkmus J E, Reardon J E, Thomson J A L 1973 Handbook of infrared radiation from combustion gases. NASA SP-3080, Washington, DC
- Magnussen B F, Hjertager B H 1976 On mathematical modeling of turbulent combustion with special emphasis on soot formation and combustion. *16th Symp. (Int.) on Combustion* (Philadelphia: The Combustion Institute) pp. 719–729
- Moult A, Spalding D B, Markatos N C G 1979 The solution of flow problems in highly irregular domains by the finite-difference method. *Trans. Inst. Chem. Eng.* 57: 200–204
- Muysenberg H P H, Simonis F, Heijden T C M van der 1993 The coupling between a glass tank and combustion chamber model. *Proc. II Int. seminar on mathematical simulation in the glass melting*, Vsetin, Czech Republic
- Patankar S V 1980 *Numerical heat transfer and fluid flow* (New York: Hemisphere)
- Post L 1987 A mathematical model of the combustion chamber in a glass furnace. *Numerical methods in thermal problems* (eds) R W Lewis, K Morgan, W G Habashi (eds) (Swansea: Pineridge) part 1
- Post L 1988 *Modeling of flow and combustion in a glass melting furnace*. Ph D thesis, Delft University of Technology, Delft
- Post L, Hoogendoorn C J 1987 Heat transfer in gas-fired glass furnaces. *Verbrennung und Feuerungen*, 13. Deutscher Flammentag, VDI-Ber. no. 645, pp. 457–466
- Siddal R G 1986 Accurate evaluation of radiative direct-exchange areas for rectangular geometries. *Proc. 8th Int. Heat Transfer Conf.*, San Francisco (New York: Hemisphere) pp. 751–755
- Simonis F, Heijden T van der, Beerkens R G C, Muijsenberg H P H 1993 Simulation of the redox distribution in glass melts and gasbubble tracing. *Proc. II Int. Sem. on mathematical simulation in the glass melting*, Vsetin, Czech Republic
- Simonis F, Waal H de, Beerkens R C G 1986 Influence of furnace design and operation parameters on the residence time distribution of glass tanks, predicted by 3-D computer simulations. *XIV International Congress on Glass*, New Delhi, collected papers, vol. 3 pp. 118–128
- Soufiani A, Hartman J, Taine J 1986 Validity of band-model calculations for  $\text{CO}_2$  and  $\text{H}_2\text{O}$  applied to radiative properties and conductive-radiative transfer. *J. Quant. Spectrosc. Radiat. Transfer* 33: 243–257
- Spalding D B 1980 Mathematical modeling of fluid-mechanics, heat-transfer and chemical reaction processes. CFDU Report HTS/80/1, Imperial College, London
- Syed K J, Stewart C D, Moss J B 1990 Modeling soot formation and thermal radiation in buoyant diffusion flames. *23rd Symp. (Int.) on Combustion* (Philadelphia: The Combustion Institute) pp. 1533–1541
- Ungan A 1985 *Three-dimensional numerical modelling of glass melting process*. Ph D thesis, Purdue University, Lafayette, Indiana
- Ungan A, Viskanta R 1987a Identification of the structure of the thermal flow in an idling container glassmelter. *Glass Technol.* 28: 252–260
- Ungan A, Viskanta R 1987b Three-dimensional numerical modeling of circulation and heat transfer in a glass melting tank. Pt. 1. Mathematical formulation, and Pt. 2. Sample simulations. *Glastech. Ber.* 60: 71–78, 115–124
- Van de Kamp W L, Swart J P, Nakamura T, Morgan M E 1991  $\text{NO}_x$  reduction and heat transfer characteristics in gas fired glass melting furnaces. *Proc. 2nd European Conference on Industrial Furnaces and Boilers*, Lisbon
- Warnatz J 1983 Hydrocarbon oxidation at high temperatures. *Ber. Bunsenges Phys. Chem.* 87: 1008–1022
- Wieringa J A 1992 *Spectral radiative heat transfer in gas fired furnaces*. Ph D thesis, Delft University of Technology, Delft
- Zel'dovich J 1946 The oxidation of nitrogen in combustion and explosions. *Acta Physicochim. URSS* 22: 577–628





## Natural convection of a radiating fluid in a square enclosure with perfectly conducting end walls

A YÜCEL, S ACHARYA\* and M L WILLIAMS

Louisiana State University, Baton Rouge, LA 70803, USA

**Abstract.** Combined natural convection and radiation in an asymmetrically heated square enclosure is studied numerically with both adiabatic and perfectly conducting end walls. The momentum and energy equations are solved by a control volume based finite difference algorithm which is coupled with the discrete ordinates method for radiative heat transfer calculations. The changes in the flow patterns and temperature distributions due to the presence of radiation in an enclosure with conducting end walls are compared with those for the case of an enclosure with adiabatic end walls, and significant differences are noted. The flow field is stronger, and the heat input along the hot wall and the end walls are greater for the conducting end wall case. The effects of optical thickness, scattering and wall emissivity on the flow and temperature fields and heat transfer rates are analysed.

**Keywords.** Natural convection; radiation; enclosure.

### 1. Introduction

Thermal radiation can strongly interact with enclosure natural convection in many situations of engineering interest such as nuclear reactor safety, combustion, fire and plumes, porous media and solar collectors. Ostrach (1988) provides a comprehensive review of studies where natural convection is the only mode of heat transfer in enclosures. Studies dealing with the interaction between radiation and natural convection have been the subject of reviews by Viskanta (1984), Yang (1986) and Viskanta & Menguc (1987). Representative studies include those reported by Lauriat (1982a, b) who used the *P*-1 differential approximation to study radiation-convection interactions of a participating medium in vertical enclosures. Desrayaud & Lauriat (1985) later extended the study for a radiating vertical fluid layer. Using the *P*-1 approach, Fusegi & Farouk (1986, pp. 81–8) studied the interaction in a square enclosure with asymmetrical heating. They later extended the study for fire-spread applications (Fusegi & Farouk 1987, pp. 63–8). Webb & Viskanta (1987) studied the

---

\*For correspondence

natural convection caused by the irradiation of a fluid layer. Yücel *et al* (1989) used the discrete ordinates method to study the combined natural convection–radiation interactions in an enclosure with a participating medium. Kassemi & Duval (1990) employed a zonal approach to analyse the effects of radiation on the transport process in rectangular enclosures. Kassemi & Naraghi (1990) used a discrete exchange factor method described by Naraghi & Kassemi (1989) to study radiation–natural convection interactions in a square enclosure. Tan & Howell (1991) used the product integration method (PIM) proposed by Tan (1989) together with a nonlinear-SOR strategy to study radiation–natural convection in a square enclosure.

In the above-mentioned studies of natural convection–radiation interactions in asymmetrically heated enclosures, adiabatic conditions are customarily assumed on the end walls. Yet, adiabatic conditions are not easily realized in experimental settings or engineering applications. The effects of conducting end walls should be taken into consideration for more realistic simulations. In this paper, two-dimensional natural convection of a radiating fluid in an externally heated vertical square enclosure is studied. The end walls are assumed to be perfectly conducting, and results are compared with results obtained with the assumption of adiabatic end walls. The perfectly conducting end wall condition has been shown to be a realistic end wall condition, with gas as the working fluid (ElSherbiny *et al* 1982) and represents a first approximation to account for the influence of conduction along the enclosure walls.

Various solution methods are currently available for the solution of radiative fluxes. These include: the zonal method, the flux method, the ray-tracing method, the Monte-Carlo method, the discrete exchange factor method, the product integration method, the finite volume method and the discrete ordinates method. More details on many of these methods can be found in Viskanta (1984), and Viskanta & Menguc (1987). In this paper, the discrete ordinates method is used to solve the radiation part of the problem. In comparison with many other competing methods, the discrete ordinates method has been shown to yield higher order accurate solutions to the equations of radiative heat transfer in problems involving radiation affected buoyant enclosure flows (Yücel *et al* 1989). In the present work, the radiation calculations using the discrete ordinates method are coupled with a control volume based finite difference algorithm which is used to solve the momentum and energy equations. The changes in the natural convection flow patterns and temperature and heat transfer distributions due to the presence of radiation are analysed and results compared for both linear temperature and adiabatic boundary conditions.

## 2. Analysis

The physical model consists of a gray, absorbing, emitting and isotropically scattering fluid in a square enclosure (figure 1). The fluid is Newtonian and incompressible; viscous dissipation is neglected and physical properties are taken as constant except for the density. The flow is assumed to be laminar, steady and two-dimensional. With these assumptions, the equations for the conservation of mass, momentum and energy can be expressed in dimensionless form as follows.

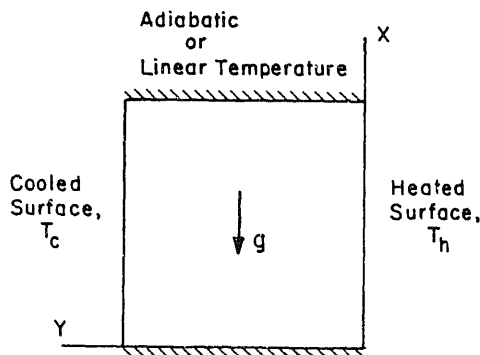


Figure 1. Schematic representation of the physical system.

$$U \frac{\partial U}{\partial X} + V \frac{\partial U}{\partial Y} = - \frac{\partial P}{\partial X} + \nabla^2 U + \frac{Ra_E}{Pr} \phi, \quad (2)$$

$$U \frac{\partial V}{\partial X} + V \frac{\partial V}{\partial Y} = - \frac{\partial P}{\partial Y} + \nabla^2 V, \quad (3)$$

$$U \frac{\partial \phi}{\partial X} + V \frac{\partial \phi}{\partial Y} = \frac{1}{Pr} \nabla^2 \phi - \frac{\phi_0}{Pl Pr} \nabla \cdot \mathbf{Q}_R, \quad (4)$$

where the Boussinesq approximation ( $\rho = \rho_0[1 - \beta(T - T_0)]$ ) was used in the buoyancy term to allow for the variation of density with temperature. In the above equations, the buoyancy induced by external heating is represented by the external Rayleigh number  $Ra_E$ .

The term  $(\phi_0/Pl \cdot Pr) \nabla \cdot \mathbf{Q}_R$  in the energy equation represents the radiative energy deposition and is determined from the solution of the radiative transfer equation. The equation governing the steady state dimensionless radiation intensity  $I$  is given by

$$\mu(\partial I / \partial X) + \xi(\partial I / \partial Y) + \tau I = (\tau/4\pi) \left[ (1 - \omega)B + \omega \int_{4\pi} I d\Omega \right], \quad (5)$$

where  $I$  is the dimensionless intensity of radiation at a point  $(X, Y)$  in the direction  $\Omega$ . The emissive power  $B$  is expressed as

$$B = (T/T_0)^4 = (\phi/\phi_0 + 1)^4. \quad (6)$$

The radiative heat flux vector can be calculated by the intensity distribution,

$$\mathbf{Q}_R = \int_{4\pi} \Omega I d\Omega = \int_{4\pi} (\mu \mathbf{i} + \xi \mathbf{j}) I d\Omega. \quad (7)$$

Integrating (5) over the whole solid angle yields a relation for the divergence of the radiative heat flux vector

$$\nabla \cdot \mathbf{Q}_R = (1 - \omega)\tau \left[ B - \int_{4\pi} I d\Omega \right], \quad (8)$$

which is substituted in the energy equation.

## 2.1 Boundary conditions

No-slip (zero velocity) boundary conditions are imposed on the rigid walls. Two end-wall boundary conditions are considered for temperature: either thermally insulated (adiabatic case) or perfectly conducting (linear temperature case). These conditions along  $X = 0$  and  $X = 1$  are expressed as:

$$\text{adiabatic case:} \quad \mathbf{Q}_T \cdot \mathbf{i} = (-\nabla \phi + (\phi_0/\text{Pl})\mathbf{Q}_R) \cdot \mathbf{i} = 0, \quad (9a)$$

$$\text{linear temperature case:} \quad \phi(X, Y) = 0.5 - Y. \quad (9b)$$

The two other walls are maintained at temperatures  $T_h$  (hot wall) and  $T_c$  (cold wall) respectively. For black surfaces the intensity of radiation leaving a bounding surface is simply the blackbody intensity of radiation emitted at the surface temperature. For diffusely reflecting and emitting end walls, the boundary intensity in the direction  $\Omega$  leaving the surface is given by

$$I(X, Y, \Omega) = \frac{\varepsilon B(X, Y)}{4\pi} + \frac{(1-\varepsilon)}{\pi} \left[ \int_{\mathbf{n} \cdot \Omega < 0} |\mathbf{n} \cdot \Omega| I(X, Y, \Omega') d\Omega' \right], \quad \text{at} \\ X = 0 \quad \text{and} \quad X = 1, \quad (10)$$

where  $\varepsilon$  is the surface emissivity ( $\varepsilon = 1$  for black surfaces) and  $\mathbf{n}$  is the unit normal at the surface.

## 2.2 Solution procedure

Equations (1)–(8) with the associated boundary conditions provide a complete mathematical formulation of the problem. The mass, momentum and energy equations are solved by a control-volume based finite difference procedure. Pressure-velocity interlinkage is resolved using the SIMPLER algorithm (Semi Implicit Method for Pressure Linked Equations Revised). This method is described in detail by Patankar (1980). The solutions to the equation of radiative transfer are obtained by the discrete ordinates method. In the discrete ordinates formulation, the solid angle is discretized by selecting an ordered set of directions with predetermined weights. Quantities involving integrals over the solid angle are evaluated by sums over the discrete directions with weights as the weighting functions. The spatial discretization introduced to obtain a finite difference form of the radiative transfer equation employs the same control volumes (cells) as the SIMPLER algorithm. For example, for a control volume bounded by  $(X_i, X_{i+1})$  and  $(Y_j, Y_{j+1})$ , the discretized form of (5) can be expressed as

$$\mu_m \frac{(I_{i+1} - I_i)}{(X_{i+1} - X_i)} + \xi_m \frac{(I_{j+1} - I_j)}{(Y_{j+1} - Y_j)} + \tau I = S, \quad (11)$$

where  $I$  is the intensity at the centre of the control volume, and  $\mu_m$  and  $\xi_m$  are the direction cosines of the vector  $\Omega$ . The above equation represents the radiative energy balance for a control volume, with the first two terms on the left-hand side of the

equation representing the loss of radiant energy across the control volume boundaries, the third term representing the loss by absorption and scattering, and the source term on the right representing the gains due to inscatter from other directions and due to emission.

The interface and cell-centre  $I$ -values are related using the linear or the step interpolation scheme which, in the  $i$ - and  $j$ -directions, are mathematically expressed as

$$I = aI_{i+1} + (1-a)I_i = aI_{j+1} + (1-a)I_j, \quad (12)$$

with  $a = 0.5$  for the linear model and  $a = 1$  in the step model. Equations (11) and (12) together lead to the discrete ordinates equation, which is solved by sweeping away from the boundaries. In the presence of scattering media or reflecting boundaries, the discrete ordinates equations are coupled and therefore have to be solved iteratively. Details of the discrete ordinates methodology can be found in Duderstadt & Martin (1979), Fiveland (1987, pp. 9–18) and Yücel *et al* (1989).

The coupled mass, momentum, energy and radiative transfer equations are solved through a global iterative procedure which is outlined below:

- (1) Initial velocity and temperature fields are assumed.
- (2) The radiative transfer equation is solved for the given temperature field. The incident radiant energy term in (8) and the boundary radiative heat flux terms in (9a) are then determined.
- (3) Equations (1)–(4) are solved to obtain a new flow field and temperature distribution.
- (4) The next iteration is performed by repeating steps 2 and 3. The procedure is continued until convergence is obtained on velocity and temperature to within a prespecified tolerance ( $|\phi_{ij}^{n+1} - \phi_{ij}^n|/|\phi_{ij, \max}^n| \leq 10^{-5}$ ).

### 2.3 Numerical uncertainty

The calculations are performed on a  $32 \times 32$  nonuniform grid packed towards the walls (generated by an algebraic stretching procedure with a packing factor of 1.5) which was found to be adequate for natural convection in externally heated vertical and inclined enclosures (Acharya & Goldstein 1985; Yücel *et al* 1989). Table 1 provides a comparison, for two grid sizes, of the values of the dimensionless total and radiative heat fluxes along the hot wall for the adiabatic end wall case. It can be seen that the two solutions agree to within 1% of each other. Furthermore, the overall energy

**Table 1.** Dimensionless average heat fluxes on the hot wall for the adiabatic end case,  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\omega = 0$ .

$\tau$	32 × 32 Grid		50 × 50 Grid	
	Total, $\bar{Q}_T$	Radiation $\bar{Q}_R \phi_0 / Pl$	Total, $\bar{Q}_T$	Radiation $\bar{Q}_R \phi_0 / Pl$
Non-radiating	13.83	0	13.76	0
5.0	31.93	23.77	31.76	23.64
1.0	39.36	31.63	39.21	31.55
0.2	46.45	37.67	46.11	37.40

balance requires

$$R = \int_{x=0}^1 [(\mathbf{Q}_T \cdot \mathbf{j})_{Y=0} - (\mathbf{Q}_T \cdot \mathbf{j})_{Y=1}] dX + \int_{x=0}^1 [(\mathbf{Q}_T \cdot \mathbf{i})_{X=0} - (\mathbf{Q}_T \cdot \mathbf{i})_{X=1}] dY = 0. \quad (13)$$

In all cases presented here,  $E$  differs from zero in only the third digit after the decimal point.

### 3. Results and discussion

In this section results are presented in the form of streamline and isotherm contour plots, midplane velocity profiles and heat transfer rates on the walls. The dimensionless stream function is obtained from the velocity field solution by evaluating the integral

$$\psi = \int_0^1 U dY, \quad (14)$$

along constant  $X$  lines with  $\psi = 0$  at  $X = Y = 0$ .

Numerical solutions are obtained for a Prandtl number of 0.72. The reference temperature ratio is taken to be  $\phi_0 = 1.5$ , i.e.  $T_c/T_h = 0.5$ . The constant fluid property assumption and the Boussinesq approximation are reasonable for the above  $T_c/T_h$  ratio (Chang *et al* 1983).

We consider a base case with  $Ra_E = 5 \times 10^6$  to demonstrate radiation affected temperature and buoyant flow fields in a square enclosure. The assumption of two-dimensional laminar flow is valid for the above value of the external Rayleigh number (Larson & Viskanta 1976; Acharya & Goldstein 1985). The isotherm and streamline contour plots for a nonradiating fluid are shown in figure 2 for the adiabatic case and in figure 3 for the linear temperature case. The temperature and flow fields are centro-symmetric in the absence of radiation. The inner core is thermally-stratified and multicellular and is surrounded by an outer convective roll rotating in an anti-

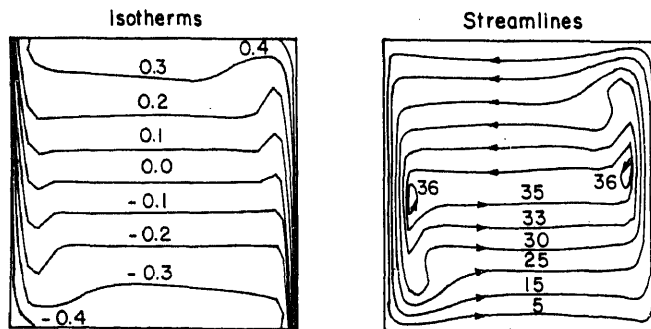
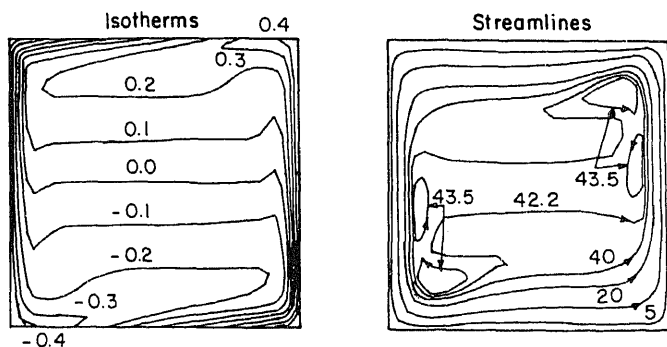


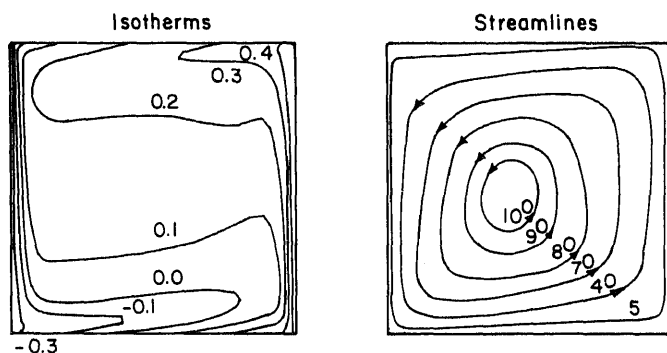
Figure 2. Isotherms and streamlines for a nonradiating fluid (adiabatic case:  $Ra_E = 5 \times 10^6$ ).



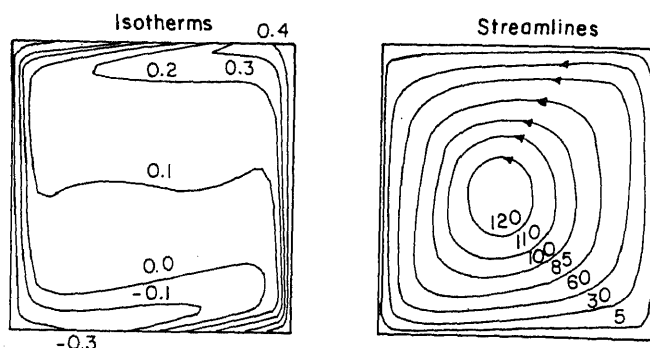
**Figure 3.** Isotherms and streamlines for a radiating fluid (linear temperature case:  $Ra_E = 5 \times 10^6$ ).

clockwise sense. Large velocity and temperature gradients indicating a boundary layer structure exist in regions adjacent to the thermally active (hot and cold) walls. For the adiabatic case, the isotherms are not orthogonal to the adiabatic walls when radiation is present. Thus, while the total flux at the wall is zero, the individual contributions due to radiation and conduction are non-zero. In the case of perfectly conducting end walls strong temperature gradients are also present near the end walls. As a result, the multicellular inner core structure extends toward the upper hot and lower cold corners of the enclosure.

Radiation effects for an absorbing, emitting but nonscattering fluid ( $\omega = 0$ ) are illustrated in figures 4 and 5. The isotherms and streamlines presented are for an optical thickness of one and a Planck number of 0.02 which entail significant heating of the fluid by radiative transfer. The enclosure walls are assumed to be black unless otherwise noted. For the case with adiabatic end walls (figure 4), the central temperature stratification is less apparent. Temperatures above the mean temperature  $T_0$  ( $\phi > 0$ ) prevail throughout most of the enclosure reflecting the penetration of the hot fluid; colder regions ( $\phi < 0$ ) are confined to the region in the vicinity of the cold wall and the lower left part of the enclosure. The flow pattern is significantly altered due to radiation, with the most significant effect being the disappearance of the



**Figure 4.** Isotherms and streamlines for a radiating fluid (adiabatic case:  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 1.0$ ,  $\omega = 0$ ).

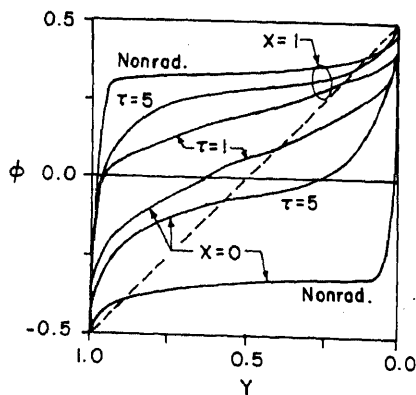


**Figure 5.** Isotherms and streamlines for a radiating fluid (linear temperature case:  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 1.0$ ,  $\omega = 0$ ).

multicellular inner core. The flow is unicellular and significantly stronger. For the perfectly conducting end walls (figure 5) the penetration of the hot wall radiation into the medium is not as large as in the adiabatic case but is still quite significant. The unicellular flow is more vigorous in this case.

Figure 6 shows the temperature variations along both end walls for the adiabatic case. For the nonradiating fluid, the temperature remains relatively uniform over the length of the walls but drops (or rises) sharply at the corners. For the radiating fluid, radiation is the dominant heat transfer mechanism. The end wall temperature variations are closer to the linear temperature condition (dashed line); as a consequence, in the presence of radiation, the temperature and flow patterns of the adiabatic and linear temperature cases (figures 4 and 5) are more similar to each other than in the non-radiating cases.

The vertical and horizontal velocity profiles along the respective horizontal and vertical midplanes are shown in figure 7. When radiation is not accounted for, the inner core is rather stagnant. Radiation causes an overall increase in the velocities and the constant-velocity-core disappears. Large velocity gradients exist along both the thermally active walls and the end walls. As noted earlier, the flow is stronger in



**Figure 6.** Temperature profiles on the end walls ( $X=0$  and  $X=1$ ) for the adiabatic case  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\omega = 0$ .



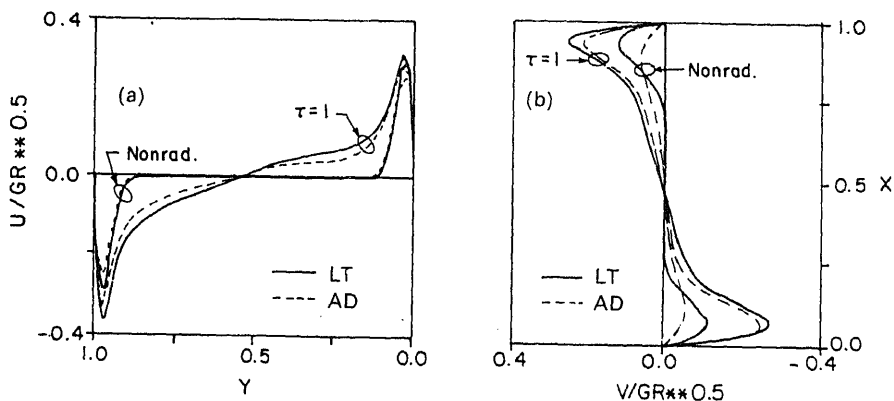


Figure 7. Velocity profiles for a radiating fluid: (a) Vertical velocity on the  $X=0.5$  midplane; (b) horizontal velocity on the  $Y=0.5$  midplane ( $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 1.0$ ,  $\omega = 0$ ; AD = adiabatic case, LT = linear temperature case).

the linear temperature case as seen by the increased magnitudes of velocities compared to the adiabatic case.

The effects of the optical thickness on the isotherm and streamline patterns are presented in figures 8 and 9 for an enclosure with perfectly conducting end walls. Figure 8 presents the temperature and flow field for an optical thickness of 5 and the general behaviour is similar to that with an optical thickness of unity. The flow is unicellular. On the other hand, for  $\tau = 0.2$ , the radiative interactions between the enclosure walls are greater in an optically thinner medium because of lessened attenuation by the fluid. The colder regions extend further into the mid-region (figure 9) and a strong core temperature stratification manifests itself. The resulting flow exhibits a multi-cellular pattern. The inner core consists of a primary convective roll in the lower half and a secondary roll in the upper half.

Figure 10 illustrates the changes in the temperature and flow fields for an isotropically scattering fluid with  $\omega = 0.5$ . Compared to the nonscattering fluid case (figure 5), there is less absorption of radiation by an absorbing and scattering fluid.

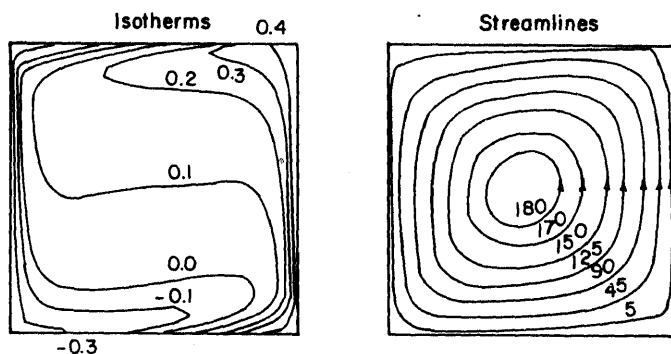
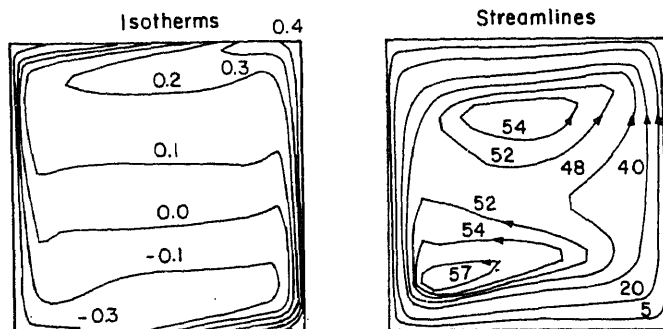


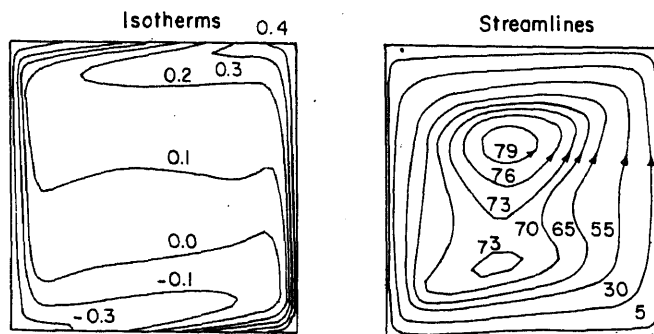
Figure 8. Isotherms and streamlines for a radiating fluid (linear temperature case:  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 5.0$ ,  $\omega = 0$ ).



**Figure 9.** Isotherms and streamlines for a radiating fluid (linear temperature case:  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 0.2$ ,  $\omega = 0$ ).

Therefore the temperature distribution is more similar to that for an optically thinner nonscattering fluid (figure 9). The flow pattern also exhibits two distinct rolls in the inner core surrounded by an outer wall-adjacent roll, all rotating in the counter-clockwise direction. Compared to the adiabatic case (not shown) where nearly symmetrical inner rolls are obtained, the inner flow pattern is less symmetrical in the conducting end wall case.

Average values of the dimensionless total heat flux (conductive and radiative combined) and the radiative contribution to the total heat flux are presented in table 2. Overall, the heat transfer rates are much higher for a radiating fluid. For  $Pl = 0.02$ , the net heat input (at  $X = 0$  and  $Y = 0$ ) is about 5–12% higher for the linear end-wall-temperature case than that for the adiabatic case. The ratio of the total and the radiative fluxes on the hot wall ( $Y = 0$ ) vary little from one case to the other. Radiation accounts for around 75–80% of the total heat flux on the hot wall in all three cases ( $\tau = 0.2$ , 1.0, 5.0). There is an additional but small heat input from the lower end wall ( $X = 0$ ) in the linear temperature case. Its magnitude increases with increasing optical thickness and the corresponding radiative contribution increases from a net heat loss (output) to a net heat gain (input). The total heat output from the cold wall ( $Y = 1$ ) is obviously identical to the total heat input from the hot wall for the adiabatic case, but is around 20% lower for the linear temperature



**Figure 10.** Isotherms and streamlines for a scattering fluid (linear temperature case:  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 1.0$ ,  $\omega = 0.5$ ).

**Table 2.** Average heat fluxes on the enclosure walls ( $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\omega = 0$ ).

$\tau$		Hot wall ( $Y = 0$ )		Cold wall ( $Y = 1$ )		Bottom end wall ( $X = 0$ )		Top end wall ( $X = 1$ )	
		$Q_{T^+}$	$(Q_R \phi_0 / Pl)^*$	$Q_T$	$(Q_R \phi_0 / Pl)$	$Q_T$	$(Q_R \phi_0 / Pl)$	$Q_T$	$(Q_R \phi_0 / Pl)$
Nonrad.	AD <sup>a</sup>	13.83	0.00	13.83	0.00	0.00	0.00	0.00	0.00
	LT <sup>b</sup>	10.69	0.00	10.69	0.00	5.28	0.00	5.28	0.00
0.2	AD	46.45	37.67	46.45	31.87	0.00	-5.09	0.00	-2.07
	LT	47.52	38.22	38.73	27.37	1.14	-3.33	9.93	4.31
1.0	AD	39.36	31.63	39.36	24.39	0.00	-3.84	0.00	-1.87
	LT	40.38	32.40	31.96	20.14	2.25	-1.24	10.67	4.70
5.0	AD	31.93	23.77	31.93	16.05	0.00	-1.72	0.00	-0.79
	LT	31.54	23.78	24.79	12.39	6.33	2.98	13.08	6.53

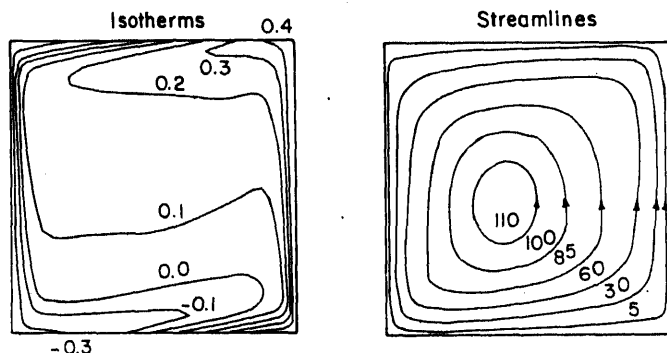
<sup>a</sup>AD: adiabatic case; <sup>b</sup>LT: linear temperature case; <sup>+</sup> dimensionless total flux; \* dimensionless radiative flux.

case. The radiative contribution to the total heat flux on the cold wall drops down to 50–65% for both cases. Approximately one-fourth of the total heat output from the enclosure (at  $X = 1$  and  $Y = 1$ ) is transferred out from the top end wall ( $X = 1$ ) for the linear temperature case; about half of this amount is by radiation.

The effect of nonblack (reflecting) end walls on the boundary layer structure along the hot and cold walls is small as can be seen in figure 11. The radiative heat fluxes on the end walls are drastically reduced for  $\varepsilon = 0.2$ . The net heat input, with a small contribution from the lower end wall, remains relatively unchanged, but the total heat output from the cold wall increases to compensate for the decreased radiative heat loss from the top end wall.

### 3.1 Computational issues

The computations were performed on a Floating Point Systems 264 Attached Processor with an IBM 3084 computer as the front end machine. A typical case with



**Figure 11.** Isotherms and streamlines for a radiating fluid (linear temperature case with end wall emissivity of 0.2:  $Ra_E = 5 \times 10^6$ ,  $Pl = 0.02$ ,  $\tau = 0.2$ ,  $\omega = 0$ ).

radiation using S4 quadrature required about 500 iterations (0.37 CPU seconds per iteration) for the  $32 \times 32$  grid. The difference in CPU times between a scattering and nonscattering case was insignificant. Although problems with scattering require inner iterations in the discrete ordinates calculations, only one inner iteration is needed per global iteration after the first 10–15 outer iterations and as the calculations move rapidly towards convergence. This trend also holds for problems with reflective boundaries which also require inner iterations.

#### 4. Conclusions

When radiation is the predominant mechanism of heat transfer, the effects of radiation on natural convection in an externally heated square enclosure with perfectly conducting end walls are similar to, but more pronounced than those in an enclosure with adiabatic end walls. The temperature distribution is significantly altered by radiation, resulting in higher temperatures in the inner region of the enclosure and, for adiabatic end walls, an end wall temperature closer to the linear temperature profile. Velocities are intensified in the presence of radiation. For the linear temperature case, the resulting flows are stronger than the corresponding flow in the adiabatic case and are generally characterized by a single prevalent convective roll. A multicellular inner core can develop for an optically thin or scattering fluid. Along the hot wall and the end walls, the heat input is larger for the conducting end wall case.

#### List of symbols

$B$	dimensionless emissive power, $(T/T_0)^4$ ;
$g$	gravitational acceleration;
$Gr$	Grashof number, $g\beta(T_h - T_c)L^3/\nu^2$ ;
$I$	dimensionless radiation intensity, $i/4\sigma T_0^4$ ;
$k$	thermal conductivity;
$L$	height and width of enclosure;
$P$	dimensionless pressure, $p^*/\rho_0(\nu/L)^2$ ;
$p, p^*$	thermodynamic and modified pressures, $p^* = p + \rho_0 g x$ ;
$Pl$	Planck number, $(k/L)/4\sigma T_0^3$ ;
$Pr$	Prandtl number, $\nu/\alpha$ ;
$\dot{q}$	volumetric heat generation rate;
$Q_R$	dimensionless radiative heat flux, $q_R/4\sigma T_0^4$ ;
$Q_T$	dimensionless combined heat flux, $q_T/[k(T_h - T_c)/L]$ ;
$Ra_E$	external Rayleigh number, $g\beta(T_h - T_c)L^3/\nu\alpha$ ;
$T$	temperature;
$T_h, T_c$	hot and cold wall temperatures;
$T_0$	reference (mean) temperature, $(T_h + T_c)/2$ ;
$U, V$	dimensionless velocities in $X$ - and $Y$ -directions, $u/(\nu/L)$ , $v/(\nu/L)$ ;
$X, Y$	dimensionless coordinates, $x/L$ , $y/L$ ;
$\alpha$	thermal diffusivity;
$\beta$	thermal expansion coefficient;

$\varepsilon$	wall emissivity;
$\kappa$	absorption coefficient;
$\mu, \xi$	direction cosines in $X$ - and $Y$ -directions;
$\nu$	kinematic viscosity;
$\rho$	density, $\rho_0[1 - \beta(T - T_0)]$ ;
$\rho_0$	reference density;
$\sigma$	scattering coefficient;
$\bar{\sigma}$	Stefan-Boltzmann constant;
$\tau$	optical thickness, $(\kappa + \sigma)L$ ;
$\phi$	dimensionless temperature, $(T - T_0)/(T_h - T_c)$ ;
$\phi_0$	reference temperature ratio, $T_0/(T_h - T_c)$ ;
$\psi$	stream function;
$\Omega$	direction vector, $\mu\mathbf{i} + \xi\mathbf{j}$ ;
$\omega$	scattering albedo;
$\nabla$	dimensionless gradient operator, $\mathbf{i}\partial/\partial X + \mathbf{j}\partial/\partial Y$ .

## References

- Acharya S, Goldstein R J 1985 Natural convection in an externally heated vertical or inclined square box containing internal energy sources. *J. Heat Transfer* 107: 855-865
- Chang L C, Yang K T, Lloyd J R 1983 Radiation-natural convection interactions in two-dimensional complex enclosures. *J. Heat Transfer* 105: 89-95
- Desrayaud G, Lauriat G 1985 Natural convection of a radiating fluid in a vertical layer. *J. Heat Transfer* 107: 710-712
- Duderstadt J J, Martin W R 1979 *Transport theory* (New York: Wiley)
- ElSherbiny S M, Hollands K G T, Raithby G D 1982 Effect of thermal boundary conditions on natural convection in vertical and inclined air layers. *J. Heat Transfer* 104: 515-520
- Fiveland W A 1987 Three-dimensional radiative heat transfer solutions by the discrete-ordinates method. *Fundamentals and applications of radiation heat transfer* (eds) A M Smith, T F Smith (New York: ASME) HTD-72
- Fusegi T, Farouk B 1986 Radiation-convection interactions in asymmetrically heated square enclosures. *Numerical methods in heat transfer* (eds) J L S Chen, K Vafai (New York: ASME) HTD-62
- Fusegi T, Farouk B 1987 Radiation-convection interactions of a non-gray gas in a square enclosure. *Heat and mass transfer in fire* (eds) A K Kulkarni, Y Jaluria (New York: ASME) HTD-73
- Kassemi M, Duval W M B 1990 Interactions of surface radiation with convection in crystal growth by vapor transport. *J. Thermophys. Heat Transfer* 4: 605-611
- Kassemi M, Naraghi M H N 1990 Analysis of radiation-natural convection interactions in 1-G and low-G environments using the discrete exchange factor method. *AIAA/ASME Thermophysics Conference*, Seattle
- Larson D W, Viskanta R 1976 Transient combined laminar free convection and radiation in a rectangular enclosure. *J. Fluid Mech.* 78: 65-85
- Lauriat G 1982a Combined radiation-convection in gray fluids enclosed in vertical cavities. *J. Heat Transfer* 104: 609
- Lauriat G 1982b Numerical study of the interaction of natural convection with radiation in nongray gases in a narrow vertical cavity. *Heat Transfer, Proc. of the 1982 Int. Conf.* (Washington, DC: Hemisphere) pp. 153-160
- Naraghi M H N, Kassemi M 1989 Radiative transfer in a rectangular enclosure; A discretized exchange factor solution. *J. Heat Transfer* 117: 1117-1124
- Ostrach S 1988 Natural convection in enclosures. *J. Heat Transfer* 110: 1175-1190
- Patankar S V 1980 *Numerical heat transfer and fluid flow* (New York: Hemisphere)

- Tan Z 1989 Radiative heat transfer in multidimensional emitting, absorbing and anisotropic scattering media – mathematical formulation and numerical method. *J. Heat Transfer* 111: 141–147
- Tan Z, Howell J R 1991 Combined radiation and natural convection in a square enclosure with participating medium. *Int. J. Heat Mass Transfer* 34: 79–97
- Viskanta R 1984 Radiation heat transfer. *Fortschritte Der Verfahrenstechnik* 22: 51–81
- Viskanta R, Menguc M P 1987 Radiation heat transfer in combustion systems. *Prog. Energy Combust. Sci.* 13: 97–160
- Webb B W, Viskanta R 1987 Radiation-induced buoyancy driven flow in rectangular enclosures: Experiment and analysis. *J. Heat Transfer* 109: 427–433
- Yang K T 1986 Numerical modeling of natural convection-radiation interactions in enclosures. *Heat Transfer. Proc. of the 1986 Int. Conf.* (eds) C L Tien, V P Carey, J K Ferrell (Washington, DC: Hemisphere) vol. 1, pp. 131–140
- Yücel A, Acharya S, Williams M L 1989 Natural convection and radiation in a square enclosure. *Numer. Heat Transfer* A15: 261–278

## **Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems**

RAM V MOHAN and KUMAR K TAMMA\*

Department of Mechanical Engineering, Institute of Technology,  
111 Church Street S.E., University of Minnesota, Minneapolis,  
MN 55455, USA

**Abstract.** Transient thermal analysis of engineering materials and structures by space discretization techniques such as the finite element method (FEM) or finite volume method (FVM) lead to a system of parabolic ordinary differential equations in time. These semidiscrete equations are traditionally solved using the generalized trapezoidal family of time integration algorithms which uses a constant single time step. This single time step is normally selected based on the stability and accuracy criteria of the time integration method employed. For long duration transient analysis and/or when severe time step restrictions as in nonlinear problems prohibit the use of taking a larger time step, a single time stepping strategy for the thermal analysis may not be optimal during the entire temporal analysis. As a consequence, an adaptive time stepping strategy which computes the time step based on the local truncation error with a good global error control may be used to obtain optimal time steps for use during the entire analysis. Such an adaptive time stepping approach is described here. Also proposed is an approach for employing combined FEM/FVM mesh partitionings to achieve numerically improved physical representations. Adaptive time stepping is employed throughout to practical linear/nonlinear transient engineering problems for studying their effectiveness in finite element and finite volume thermal analysis simulations.

**Keywords.** Transient thermal analysis; finite element method; finite volume method; temporal analysis.

### **Introduction**

The complexity of modern engineering systems places increased demand on thermal analysis of engineering systems with accurate physical interpretations. More and more discrete numerical methods such as the finite difference method (FDM) (Lax &

---

\*For correspondence

Wendroff 1964; Richtmeyer & Morton 1967), finite element method (FEM) (Zienkiewicz & Cheung 1965; Bathe 1982; Hughes 1987) etc., are used in conduction heat transfer analysis. The finite element method provides a close approximation of curved boundaries, and a systematic and general way of modelling the boundary conditions. For these reasons, FEM has become a powerful numerical tool in real engineering problems.

The finite difference method is another numerical technique in which the approximate solution is obtained by directly discretizing the governing differential equation. This form of the finite difference method has been historically restricted to rectangular domains and meshes and later been extended to include applicability to general, orthogonal curvilinear coordinate systems. Another approach that has received attention among heat transfer analysts and researchers is the control volume approach (Patankar & Baliga 1978; Patankar 1980; Baliga & Patankar 1983). In this approach, an energy balance is applied to discrete control volumes and discretization is used only where surface fluxes require approximation. The property of the control volume approach is that the resulting finite difference equations are conservative, and the discrete equations maintain an accurate accounting of energy flows in the domain. In the application of the finite difference and control volume methods, the coordinate system must be defined over the entire solution domain prior to effecting the discrete method. The finite element method, however, removes the above disadvantage by utilizing a coordinate system which is local to each individual element. Another approach is the use of finite element philosophy (such as isoparametric formulations based on a finite volume) directly to the conservative energy form of the heat conduction equation, and obtaining the discrete form based on the energy balance. This approach called the finite volume approach, is cited to have the benefits of the finite element method in its applicability to a general curvilinear domain, while preserving the conservation of energy (Schneider 1982; Baliga & Patankar 1983; Schneider & Zedan 1983; Banaszek 1984).

The finite element method follows a philosophically different approach than does the finite difference method. In elasticity problems, for e.g., there exists a variational extremum principle such that the minimization of potential energy or the application of the principle of virtual work leads, naturally, to the discrete model. In heat conduction, such a natural formulation with a clear physical interpretation does not exist (Zienkiewicz & Cheung 1965; Bathe 1982; Owen & Damjanic 1983; Taylor *et al* 1983, pp. 405–31; Hughes 1987; Tamma & Namburu 1989; Namburu & Tamma 1991), although the existence of variational forms for thermal analysis situations can be proven.

Traditionally, the semi-discrete time dependent equations obtained for the transient analysis are solved with a time stepping scheme using the trapezoidal family of the algorithms (Bathe 1982; Hughes 1987; Tamma & Namburu 1989; Namburu & Tamma 1991). Normally a single time step is employed over the entire temporal analysis. This single time step is normally selected based on the stability and accuracy criteria of the time integration scheme used. For a particular trapezoidal family of algorithms, the time step is based on the accuracy desired. Explicit and implicit methods have been used in the past. Whereas explicit methods are easy to code, the severe time step restriction brought by stability considerations have made implicit methods, in particular the trapezoidal Crank–Nicolson method, a logical choice by many analysts. However, although the method is unconditionally stable, the time step is still dictated by accuracy considerations. Nonetheless, the single time step



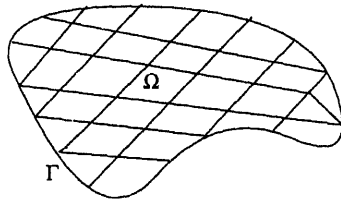


Figure 1. Typical solution domain.

selected initially for desired accuracy levels may not be optimal at all times during the analysis. It is possible that the time step selected may be too large for the accuracy desired in the analysis. It is also possible that the initial time step selected is too small during the analysis and the same desired accuracy could have been obtained using a larger time step during the analysis. An adaptive automated time stepping approach based on local truncation error with good control of global error will yield optimal time steps which changes continually during the analysis. Such an adaptive time stepping strategy based on local truncation error is considered here for the practical applicability to engineering problems in conjunction with finite element and finite volume methods. Conduction/convection/radiation effects are included. A brief overview of finite element/finite volume methods is first described followed by an effective automated time stepping approach adopted in this paper. Combined mesh partitionings involving FEM/FVM meshes based on physical situations to obtain numerically improved physical representations have been considered. Numerical test cases are described and comparative pros and cons are identified for practical situations.

### Geometry and element definition

For purposes of illustration, attention is restricted here to two-dimensional problems. A typical solution domain is shown in figure 1 which is subdivided into linear quadrilateral finite elements. A typical element is shown in figure 2, with the local coordinate system  $(\xi, \eta)$ . For the individual element, the local node numbers ranging from 1 through 4 are shown and the temperature field  $T$  and the global coordinates

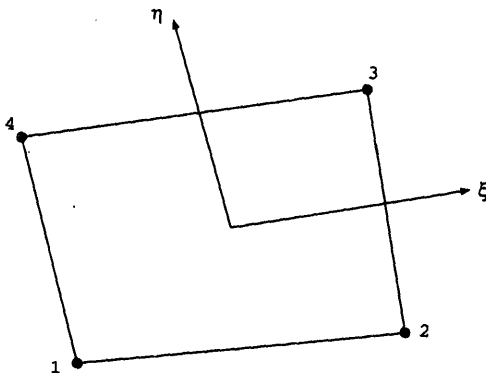


Figure 2. A typical finite volume element.

$x, y$  are interpolated using bilinear isoparametric formulations of the form

$$T = \sum_{i=1}^4 N_i T_i, \quad (1)$$

$$x = \sum_{i=1}^4 N_i x_i, \quad (2)$$

$$y = \sum_{i=1}^4 N_i y_i, \quad (3)$$

where the shape functions  $N_i$  are defined by

$$\begin{aligned} N_1 &= \frac{1}{4}(1 - \xi)(1 - \eta), \\ N_2 &= \frac{1}{4}(1 - \eta)(1 + \xi), \\ N_3 &= \frac{1}{4}(1 + \xi)(1 + \eta), \\ N_4 &= \frac{1}{4}(1 + \eta)(1 - \xi). \end{aligned} \quad (4)$$

The  $x$  and  $y$  derivatives of the temperature can be determined as

$$B = \left\{ \begin{aligned} \frac{\partial T}{\partial x} &= \sum_{i=1}^4 \frac{\partial N_i}{\partial x} T_i \\ \frac{\partial T}{\partial y} &= \sum_{i=1}^4 \frac{\partial N_i}{\partial y} T_i \end{aligned} \right\}, \quad (5)$$

where the  $x$  and  $y$  derivatives of the shape functions appearing in (5) are determined by

$$\begin{Bmatrix} \frac{\partial N_i}{\partial x} \\ \frac{\partial N_i}{\partial y} \end{Bmatrix} = \frac{1}{\text{Det}[J]} \begin{bmatrix} \frac{\partial y}{\partial \eta} & -\frac{\partial y}{\partial \xi} & \frac{\partial N_i}{\partial \xi} \\ -\frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \xi} & \frac{\partial N_i}{\partial \eta} \end{bmatrix}. \quad (6)$$

Considering a general line segment shown in figure 3, the normal vector while traversing along from point 1 to point 2 is defined by

$$d\mathbf{S} = dy\hat{i} - dx\hat{j}. \quad (7)$$

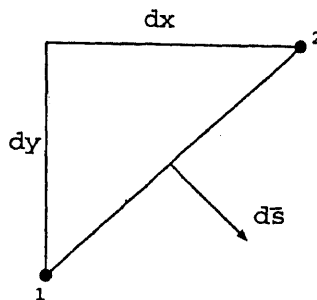


Figure 3. Typical surface line segment.

The above definitions set the basis in the development of the finite element/finite volume formulations as described next.

### Finite element equations and element matrices

Of interest here are general nonlinear transient thermal problems governed by

$$\rho_c(\partial T/\partial t) - \nabla \cdot (K \nabla T) = Q, \quad (8)$$

with the appropriate boundary and initial conditions. These are defined as

$$T = T_p, \quad \text{on } \Gamma_1, \quad (9)$$

$$q_i n_i - q_s + q_h + q_r = 0, \quad \text{on } \Gamma_2, \quad (10)$$

where

$$q_h = h(T - T_h), \quad (11)$$

$$q_r = \sigma \varepsilon (T^4 - T_r^4). \quad (12)$$

Applying the traditional Galerkin finite element method to (8–12), using the shape functions defined in the earlier section, leads to a semi-discrete equation of the form

$$[C]\{\dot{T}\} + [[K_c] + [K_h] + [K_r]]\{T\} = \{R_c\} + \{R_Q\} + \{R_q\} + \{R_h\} + \{R_r\} \quad (13)$$

where  $[C]$  is the capacitance matrix,  $[K_c]$ ,  $[K_h]$  and  $[K_r]$  are element conduction matrices corresponding to conduction, convection and radiation respectively;  $\{R_c\}$ ,  $\{R_Q\}$ ,  $\{R_q\}$ ,  $\{R_h\}$ ,  $\{R_r\}$  are the load vectors from specified nodal temperatures, internal heat generation, edge heating, radiation and convection. These are defined as

$$[C] = \int_{\Omega_e} \rho c N^T N d\Omega, \quad (14)$$

$$[K_c] = \int_{\Omega_e} B^T K B d\Omega, \quad (15)$$

$$[K_h] = \int_{\Gamma_2} h N^T N d\Omega, \quad (16)$$

$$[K_r]\{T\} = \int_{\Gamma_2} \sigma \varepsilon T^4 N^T d\Omega, \quad (17)$$

$$\{R_c\} = - \int_{\Gamma_1} (\mathbf{q} \cdot \mathbf{n}) N^T d\Gamma, \quad (18)$$

$$\{R_Q\} = \int_{\Omega_e} Q N^T d\Omega, \quad (19)$$

$$\{R_q\} = \int_{\Gamma_2} q_s N^T d\Gamma, \quad (20)$$

$$\{R_h\} = \int_{\Gamma_2} h T_h N^T d\Gamma, \quad (21)$$

$$\{R_r\} = \int_{\Gamma_2} \sigma \varepsilon T_r^4 N^T d\Gamma, \quad (22)$$

with initial conditions  $\mathbf{T}(0) = \mathbf{T}_0$ .

### Finite volume equations and element matrices

Here a procedure analogous to that followed by the conventional finite element formulation is considered. In this a single, isolated finite volume is first considered. The application of an energy balance to this element gives rise to element level matrices, which are related to the nodal temperature values of the element. Once the appropriate capacitance and conduction and load formulations are obtained, assembly rules in a similar sense of the finite element formulations can be employed to construct the global equation system from the element level equations.

The element matrices are constructed using a single element as shown in figure 4. The single finite volume element is subdivided into four internal finite volumes, each of which is associated with the corresponding nearest neighboring node of the element. In the linear quadrilateral element shown in figure 4, the control volume boundaries are chosen to be coincident with the element exterior boundaries and with the local coordinate surfaces defined by  $\xi = 0$  and  $\eta = 0$ . This choice is consistent from element to element in the entire formulation, and this boundary selection makes the evaluation of the integrals defined in the formulation easier.

The energy balance is now applied to one such control volume and is expressed as: net rate of conduction into control volume = rate of generation within control volume + rate of change of energy within the control volume.

For the control volume associated with node 3 (sub-control volume 3) in figure 4, this energy balance can be mathematically expressed as

$$Q_{2,3} + Q_{4,3} + Q_{e1,3} + Q_{e2,3} + \iint_{CV} Q dV = \frac{\partial}{\partial t} \iint_{CV} \rho c T dV, \quad (23)$$

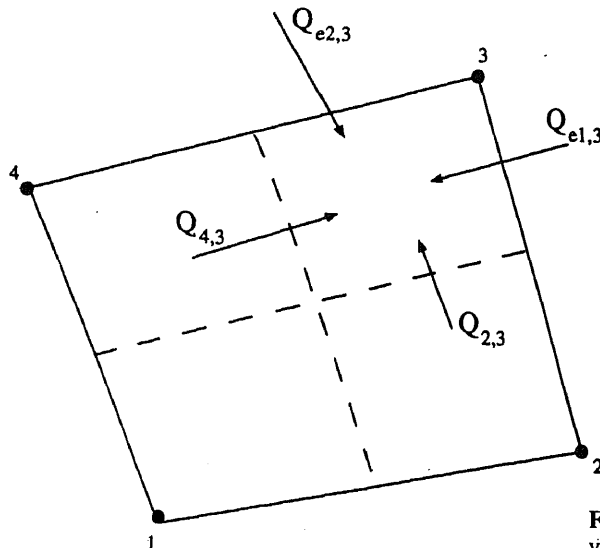


Figure 4. Single element with finite volume subdivision and heat flows.

where the limits of the integral correspond to the control volume associated with the node 3. The subscripts  $e1$  and  $e2$  in (23) refer to the energy flows into control volume 3 through surfaces which are on the exterior of the element and arise either from the physical domain boundary or from adjacent elements. In the case of adjacent elements, the heat fluxes cancel with each other, while in the case of the boundary, the boundary conditions determine the contribution of these terms. The interior terms which give rise to the conduction matrix coefficients are now considered.

In general, the heat flow through a surface can be expressed as

$$Q = \int_S \mathbf{q} \cdot d\mathbf{S} \quad (24)$$

where the heat flux vector  $\mathbf{q}$  is given as

$$\mathbf{q} = -k_x \frac{\partial T}{\partial x} \hat{i} - k_y \frac{\partial T}{\partial y} \hat{j}. \quad (25)$$

In this case

$$d\mathbf{S} = dy\hat{i} - dx\hat{j}. \quad (26)$$

Based on this the various matrices for the finite volume element can be obtained. The conduction contributions are

$$K_{1,i} = \int_{\xi=-1}^{\xi=0} \left( k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \xi} - k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \xi} \right) \bigg|_{\eta=0} d\xi + \int_{\eta=-1}^{\eta=0} \left( -k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \eta} + k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \eta} \right) \bigg|_{\xi=0} d\eta, \quad (27)$$

$$K_{2,i} = \int_{\xi=1}^{\xi=0} \left( k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \xi} - k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \xi} \right) \bigg|_{\eta=0} d\xi + \int_{\eta=-1}^{\eta=0} \left( k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \eta} - k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \eta} \right) \bigg|_{\xi=0} d\eta, \quad (28)$$

$$K_{3,i} = \int_{\eta=0}^{\eta=1} \left( k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \eta} - k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \eta} \right) \bigg|_{\xi=0} d\eta + \int_{\xi=0}^{\xi=1} \left( -k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \xi} + k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \xi} \right) \bigg|_{\eta=0} d\xi, \quad (29)$$

$$K_{4,i} = \int_{\eta=0}^{\eta=1} \left( -k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \eta} + k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \eta} \right) \bigg|_{\xi=0} d\eta + \int_{\xi=-1}^{\xi=0} \left( -k_x \frac{\partial N_i}{\partial x} \frac{\partial y}{\partial \xi} + k_y \frac{\partial N_i}{\partial y} \frac{\partial x}{\partial \xi} \right) \bigg|_{\eta=0} d\xi, \quad (30)$$

where  $i = 1, 4$  in the above expressions are for a bilinear element.

Based on the energy balances of each of the control volumes, the capacitance matrix entries are given by the following expressions

$$C_{1,i} = \int_{\xi=-1}^{\xi=0} \int_{\eta=-1}^{\eta=0} \rho c N_i \text{Det}[J] d\xi d\eta, \quad (31)$$

$$C_{2,i} = \int_{\xi=0}^{\xi=1} \int_{\eta=-1}^{\eta=0} \rho c N_i \text{Det}[J] d\xi d\eta, \quad (32)$$

$$C_{3,i} = \int_{\xi=0}^{\xi=1} \int_{\eta=0}^{\eta=1} \rho c N_i \text{Det}[J] d\xi d\eta, \quad (33)$$

$$C_{4,i} = \int_{\xi=-1}^{\xi=0} \int_{\eta=0}^{\eta=1} \rho c N_i \text{Det}[J] d\xi d\eta. \quad (34)$$

The stiffness matrix corresponding to surface convection is given by

$$(K_h)_{1,i} = \int_{\xi=-1}^{\xi=0} \int_{\eta=-1}^{\eta=0} h N_i \text{Det}[J] d\xi d\eta, \quad (35)$$

$$(K_h)_{2,i} = \int_{\xi=0}^{\xi=1} \int_{\eta=-1}^{\eta=0} h N_i \text{Det}[J] d\xi d\eta, \quad (36)$$

$$(K_h)_{3,i} = \int_{\xi=0}^{\xi=1} \int_{\eta=0}^{\eta=1} h N_i \text{Det}[J] d\xi d\eta, \quad (37)$$

$$(K_h)_{4,i} = \int_{\xi=-1}^{\xi=0} \int_{\eta=0}^{\eta=1} h N_i \text{Det}[J] d\xi d\eta. \quad (38)$$

The stiffness matrices corresponding to the surface radiation involved in the Jacobian for the nonlinear formulations are given by

$$(K_R)_{1,i} = \int_{\xi=-1}^{\xi=0} \int_{\eta=-1}^{\eta=0} 4\sigma\epsilon T_{\text{ave}}^3 N_i \text{Det}[J] d\xi d\eta, \quad (39)$$

$$(K_R)_{2,i} = \int_{\xi=0}^{\xi=1} \int_{\eta=-1}^{\eta=0} 4\sigma\epsilon T_{\text{ave}}^3 N_i \text{Det}[J] d\xi d\eta, \quad (40)$$

$$(K_R)_{3,i} = \int_{\xi=0}^{\xi=1} \int_{\eta=0}^{\eta=1} 4\sigma\epsilon T_{\text{ave}}^3 N_i \text{Det}[J] d\xi d\eta, \quad (41)$$

$$(K_R)_{4,i} = \int_{\xi=-1}^{\xi=0} \int_{\eta=0}^{\eta=1} 4\sigma\epsilon T_{\text{ave}}^3 N_i \text{Det}[J] d\xi d\eta. \quad (42)$$

The various volume load vectors for volume heating, convection and radiation are given by the following integral expressions.

For the volumetric heat source  $Q$ , the corresponding load vectors are given by

$$(R_Q)_1 = \int_{\xi=-1}^{\xi=0} \int_{\eta=-1}^{\eta=0} Q \text{Det}[J] d\xi d\eta, \quad (43)$$

$$(R_Q)_2 = \int_{\xi=0}^{\xi=1} \int_{\eta=-1}^{\eta=0} Q \text{Det}[J] d\xi d\eta, \quad (44)$$

$$(R_Q)_3 = \int_{\xi=0}^{\xi=1} \int_{\eta=0}^{\eta=1} Q \text{Det}[J] d\xi d\eta, \quad (45)$$

$$(R_Q)_4 = \int_{\xi=-1}^{\xi=0} \int_{\eta=0}^{\eta=1} Q \text{Det}[J] d\xi d\eta. \quad (46)$$

The load vectors due to surface convection are given by

$$(R_h)_1 = \int_{\xi=-1}^{\xi=0} \int_{\eta=-1}^{\eta=0} h T_h \text{Det}[J] d\xi d\eta, \quad (47)$$

$$(R_h)_2 = \int_{\xi=0}^{\xi=1} \int_{\eta=-1}^{\eta=0} h T_h \text{Det}[J] d\xi d\eta, \quad (48)$$

$$(R_h)_3 = \int_{\xi=0}^{\xi=1} \int_{\eta=0}^{\eta=1} h T_h \text{Det}[J] d\xi d\eta, \quad (49)$$

$$(R_h)_4 = \int_{\xi=-1}^{\xi=0} \int_{\eta=0}^{\eta=1} h T_h \text{Det}[J] d\xi d\eta, \quad (50)$$

where  $T_h$  is the ambient temperature. The load vectors due to surface radiation are given by

$$(R_r)_1 = \int_{\xi=-1}^{\xi=0} \int_{\eta=-1}^{\eta=0} \sigma \varepsilon T_h^4 \text{Det}[J] d\xi d\eta, \quad (51)$$

$$(R_r)_2 = \int_{\xi=0}^{\xi=1} \int_{\eta=-1}^{\eta=0} \sigma \varepsilon T_h^4 \text{Det}[J] d\xi d\eta, \quad (52)$$

$$(R_r)_3 = \int_{\xi=0}^{\xi=1} \int_{\eta=0}^{\eta=1} \sigma \varepsilon T_h^4 \text{Det}[J] d\xi d\eta, \quad (53)$$

$$(R_r)_4 = \int_{\xi=-1}^{\xi=0} \int_{\eta=0}^{\eta=1} \sigma \varepsilon T_h^4 \text{Det}[J] d\xi d\eta. \quad (54)$$

The element matrices due to edge heating, edge convection and edge radiation are given by the line integrals. For example for the control volume 3 it is given by an integral of the form

$$(K_{\text{edge}})_{3,i} = \int_{\eta=0}^{\eta=1} N_i(1, \eta) \text{Det}[S] d\eta, \quad (55)$$

where  $\text{Det}[S]$  is the determinant associated with the edge integral. For a typical edge of length  $l$  the associated stiffness matrix form is given by

$$K_{\text{edge}} = \frac{l}{2} \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}. \quad (56)$$

The load vector contribution due to the heating, radiation and convection effects on an edge is given by a vector of the form

$$R_{\text{edge}} = \frac{l}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \quad (57)$$

### Adaptive time integration methods

An automated adaptive time stepping approach is employed here for both the finite element method (FEM) and the finite volume method. This adaptive time stepping approach based on the localized error control which also gives a good global error control has been used in conjunction with the generalized trapezoidal  $\alpha$ -family of time integration methods (Bathe 1982; Hughes 1987) used for the semi-discrete heat equation. For completeness, the generalized trapezoidal family of algorithms is briefly described followed by the adaptive time stepping strategy.

### Generalized trapezoidal family method

The semidiscrete heat equation can be written as

$$C\dot{\mathbf{T}} + \mathbf{K}\mathbf{T} = \mathbf{F}, \quad (58)$$

where  $\mathbf{C}$  is the capacitance matrix,  $\mathbf{K}$  is the conduction matrix,  $\mathbf{F}$  is the heat supply vector,  $\mathbf{T}$  is the temperature vector, and  $\dot{\mathbf{T}}$  is the time derivative of  $\mathbf{T}$ . The matrices  $\mathbf{C}$  and  $\mathbf{K}$  are assumed to be symmetric. The heat supply is a prescribed function of  $t$  and  $\mathbf{T}$  comes from the various load vector terms considered earlier.

The initial-value problem consists of finding a function  $\mathbf{T}$  satisfying (58) and

$$\mathbf{T}(0) = \mathbf{T}_0, \quad (59)$$

where  $\mathbf{T}_0$  is initially given.

The most well known and commonly used algorithms for solving (58) are members of the generalized trapezoidal family of methods, which consists of the following equations

$$C\dot{\mathbf{T}}_{n+1} + \mathbf{K}\mathbf{T}_{n+1} = \mathbf{F}_{n+1}, \quad (60)$$

$$\mathbf{T}_{n+1} = \mathbf{T}_n + \Delta t \dot{\mathbf{T}}_{n+\alpha}, \quad (61)$$

$$\dot{\mathbf{T}}_{n+\alpha} = (1 - \alpha)\dot{\mathbf{T}}_n + \alpha\dot{\mathbf{T}}_{n+1}, \quad (62)$$

where  $\mathbf{T}_n$  and  $\dot{\mathbf{T}}_n$  are the approximations to  $\mathbf{T}(t_n)$  and  $\dot{\mathbf{T}}(t_n)$  respectively;  $\mathbf{F}_{n+1} = \mathbf{F}(t_{n+1})$ ;  $\Delta t$  is the time step, assumed constant for the time being; and  $\alpha$  is the parameter in the interval  $[0, 1]$  where  $\alpha = 0$  corresponds to the forward Euler scheme;  $\alpha = 1/2$  is the Crank–Nicholson scheme;  $\alpha = 2/3$  is the Galerkin scheme and  $\alpha = 1$  is the backward Euler scheme. Of these schemes, only the  $\alpha = 1/2$  is second order accurate in time.

Customarily most of the approaches adopted for the time integration have been based on a single time step being used in the entire analysis. This time step is normally selected based on the stability and accuracy criteria which depends on the value of  $\alpha$ . In particular, in the finite volume method the value of  $\alpha = 1$  has been extensively used which corresponds to the backward Euler method which is unconditionally stable, but only first order accurate. In the finite element method, however,  $\alpha = 1/2$  and  $\alpha = 0$  have been the more prominent approaches for transient thermal analysis simulations.



### Adaptive time stepping strategy

The finite element/finite volume methods traditionally employ a single time stepping strategy in conjunction with the selected generalized trapezoidal family of schemes described earlier. In this case the time step  $\Delta t$  selected for the time stepping is based on the accuracy desired. Instead of a fixed time step it is desirable to have an adaptive automated time step which varies continually during the analysis minimizing the local truncation error and with a good global error control. Such an approach will give optimal time steps with a good error control. Eriksson and Johnson (Eriksson & Johnson 1987; Johnson 1988; Thomea *et al* 1990) developed error bounds for a class of problems defined by the first-order parabolic differential equations to which the semi-discrete form of the heat equation belongs. The error bound can be written (with particular reference to the backward Euler) as

$$\max_{t \leq t_N} \| \mathbf{T}(t_n) - \mathbf{T}_n \| \leq C [\log(t_N/\Delta t_n) + 1]^{1/2} \max_{n \leq N} (\Delta t_n \max_{t \in I_n} \| \dot{\mathbf{T}}(t) \|). \quad (63)$$

The adaptive time stepping approach is based on the following strategy. Suppose,  $\delta$  is a given tolerance and the discrete equation for  $\mathbf{T}$  satisfies

$$\max_{t \leq t_N} \| \mathbf{T}(t) - \mathbf{T}_n \| \leq \delta. \quad (64)$$

Using (63) and neglecting the logarithmic error, the time step  $\Delta t_n$  can be chosen such that

$$C \Delta t_n \max_{t \leq t_N} \| \dot{\mathbf{T}}(t) \| \leq \delta, \quad (65)$$

where the constant  $C$  is known approximately. Since  $\mathbf{T}(t)$  is not known, the condition given by (65) is replaced by

$$\| \mathbf{T}_n - \mathbf{T}_{n-1} \| \leq \delta/C = \Delta T_{tol}. \quad (66)$$

This leads to an algorithm for  $\Delta t_n$ , assuming  $\mathbf{T}_{n-1}$  has already been computed. This permits an adaptive time stepping strategy as follows:

- (1) choose the initial step size  $\Delta t_0$  ( $\Delta t_{n-1}$ );
- (2) increment in time with  $\Delta t_n = \Delta t_{n-1}$  to obtain the solution  $\mathbf{T}_n$ ;
- (3) if  $\delta/\gamma C \leq \| \mathbf{T}_n - \mathbf{T}_{n-1} \| \leq \delta/C$ ;
- (4) then stop and accept the time step  $\Delta t_n$ . Otherwise increase or decrease  $\Delta t_n$  by a factor, say, 2 (for example).

Here  $\delta$  is the required selected tolerance,  $C$  is an assumed known constant,  $\gamma$  is a suitable constant (2 or 3).  $\| \cdots \|$  refers to the  $L_\infty$  norm. This gives a simple error estimator with a good control of global error.

The estimated error employing the above is thus

$$\max_{t \leq t_N} \| \mathbf{T}_{exact}(t) - \mathbf{T}(t) \| \leq \delta/C = \Delta T_{tol}. \quad (67)$$

That is,

$$\|T_{exact}(t) - \mathbf{T}(t)\| \leq \delta. \quad (68)$$

The above adaptive time stepping strategy is theoretically justified for nonlinear parabolic problems (Eriksson & Johnson 1987; Thomea *et al* 1990) and can be used in heat transfer situations and solidification problems (Ouyang & Tamma 1992). The additional computations involved in this time step control are considerably small. Limitations on permissible time steps are based on stability for the explicit form of the time integration and based on accuracy for the implicit form of the time integration.

For the adaptive time stepping strategy,  $\Delta T_{tot}$  can also be automatically selected to control the error employing the following proposed procedure. At step  $n$ .

- Set  $\Delta T_{tot} = x\%$  of  $T_r$  where  $T_r = |\mathbf{T}_{max} - \mathbf{T}_{min}|$ .

This adaptive time stepping strategy based on *a posteriori* error estimates, is reliable and for  $\|\mathbf{T}_{exact} - \mathbf{T}\| < \delta$  the error is bounded. The above-mentioned approach enables control of error and permits an effective and optimal strategy for the numerical solution of both linear and nonlinear heat transfer problems, employing finite element and finite volume approaches.

The time step selected during the time advancement based on the above scheme has been used in test problems involving only finite element/finite volume element formulations for general heat transfer problems. Comparisons are drawn for the pros/cons over single time stepping strategies.

For the adaptive time stepping strategy, additional computations at intermediate levels are necessary whenever there is a change in the time step. Caution should be exercised to adjust the time steps when they become too small or too large, and when there are rapid transient fluctuations.

### Test problems

To validate the present developments the following test problems are considered. These test problems with various material nonlinearities and radiation effects cover a wide range of potential problems encountered in engineering practice. For the transient thermal analysis the time integration has been performed using the generalized trapezoidal  $\alpha$  family of algorithms, in conjunction with both single and adaptive time stepping strategies. Finite element and finite volume based elements are used as appropriate and an illustrative mesh partitioning example is also described. All computations have been performed on Cray XMP. The test examples are described next.

### Transient nonlinear analysis of the space shuttle thermal protection system

The problem involves one-dimensional representations of the thermal protection system (TPS) of the space shuttle which is composed of different materials (figure 5). The thermophysical properties of various materials are given in figure 6. The shuttle TPS, which is initially at  $T = 322$  K is subjected to a sudden step heat flow input at the exposed surface as shown in Figure 5. The exposed surface of the TPS is assumed

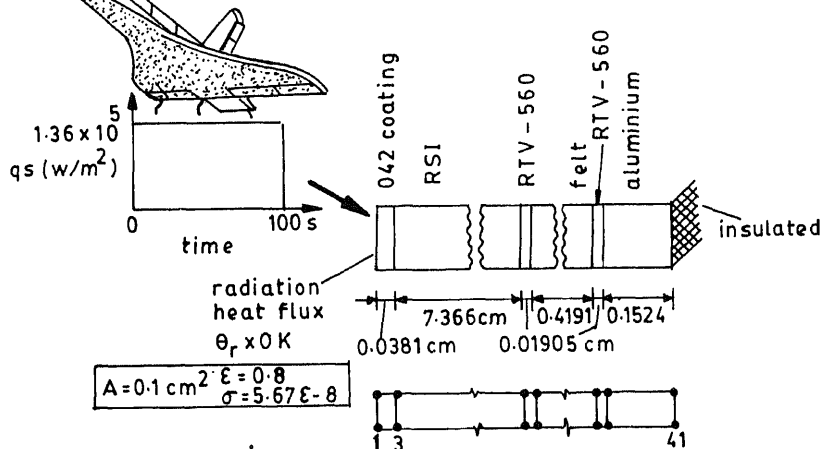


Figure 5. Shuttle TPS and description of the problem.

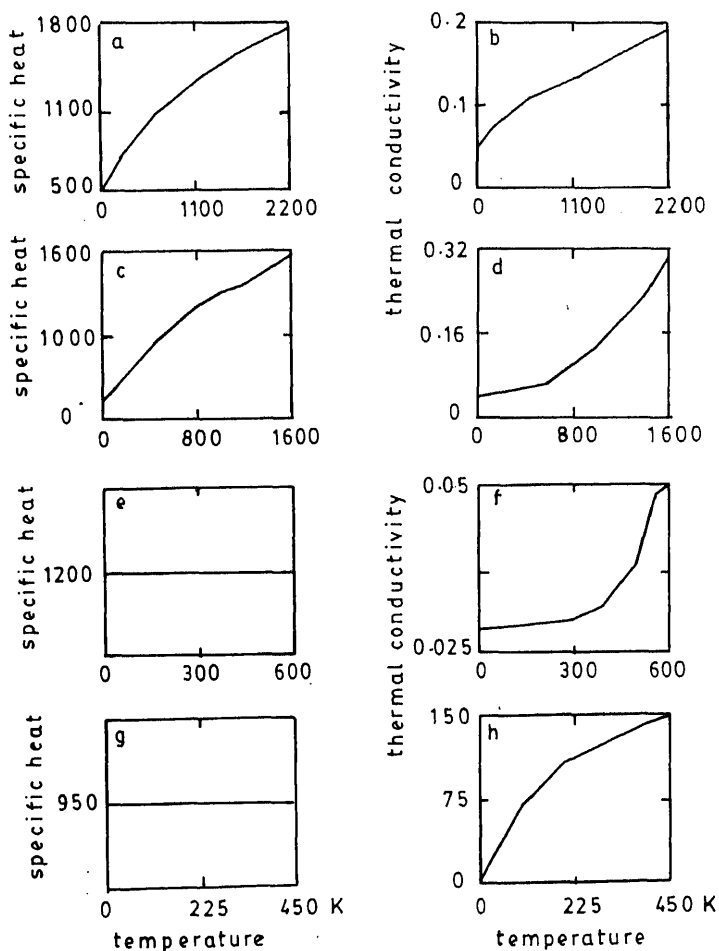


Figure 6. Thermal properties of various materials.  $C_p$  and  $K$  respectively for 042 (a & b), RSI (c & d), felt (e & f), and aluminum (g & h).

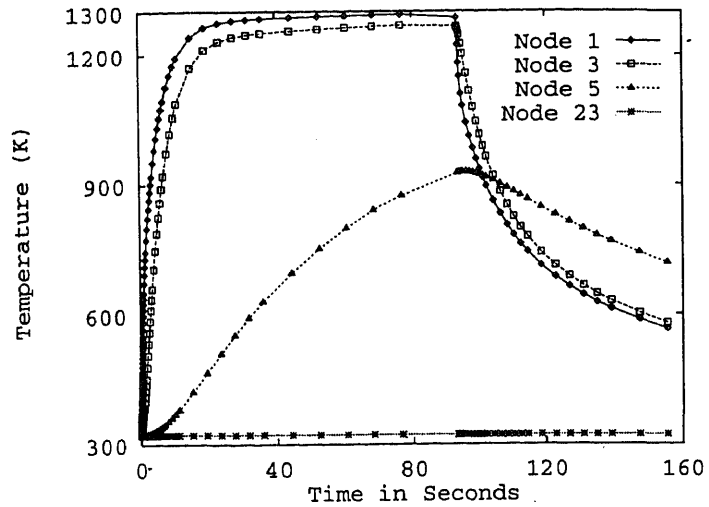


Figure 7. Temperature histories of various locations of TPS.

to radiate to an ambient absolute zero temperature continuously. The same problem was also analysed by Williams & Curry (1977), and Namburu & Tamma (1991). For evaluating the effectiveness of the finite volume formulations, a total of 19 finite volume elements with a total of 40 d.o.f. has been employed. A typical transient thermal response is shown in figure 7. The results obtained employing adaptive time stepping and without adaptive time stepping and using the Galerkin finite elements and finite volume elements show good comparison. The adaptive time stepping variations during the analysis for both the Galerkin finite element method and finite volume method are shown in figure 8. It is clear from the figure that the time step

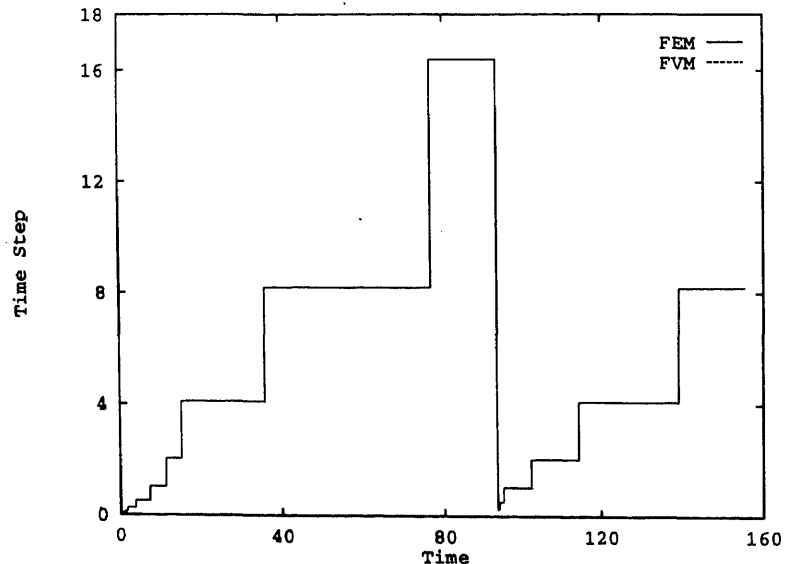
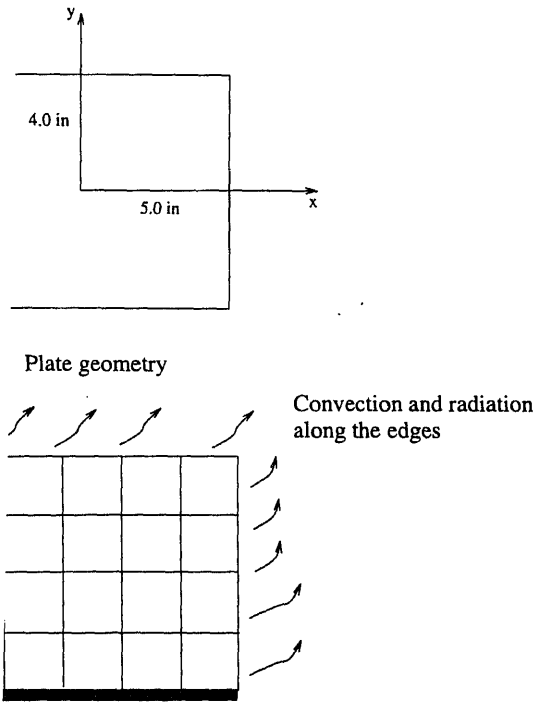


Figure 8. Adaptive time step variations.

l based on a *a posteriori* error estimator and the adaptivity is clearly seen are sudden changes in the external heat fluxes etc., as seen when the is 100 seconds. A time step of 0.01 seconds ( $\alpha = 0.5$ , extremely conservative) ed to serve as a benchmark result without adaptive time stepping. With e time stepping procedure employed, the analysis was completed in ely 200 steps for the same time period and the results are in excellent

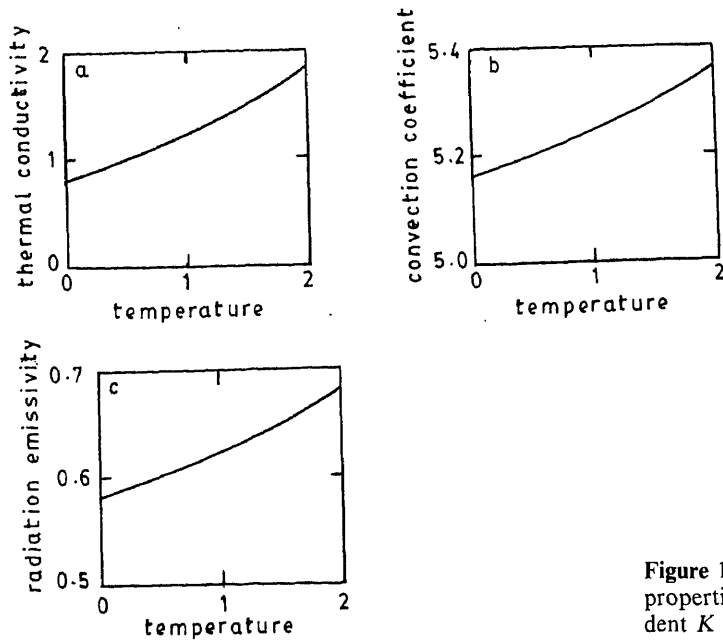
**sional rectangular plate with convection and radiation along two edges**

mple a rectangular plate with nonlinear boundary conditions involving nd nonlinear material properties is considered. One quarter of the plate l due to symmetry. The physical plate and the finite element model are figure 9. The nonlinear material properties are as shown in figure 10. n of results with traditional finite element formulation shows the effective- finite volume method as shown in figure 11. A time step of 0.05 seconds employed when the analysis is done without resorting to adaptive time hich requires 200 steps, and with adaptive time stepping the analysis is in about 120 steps. The adaptive time step variations are shown in

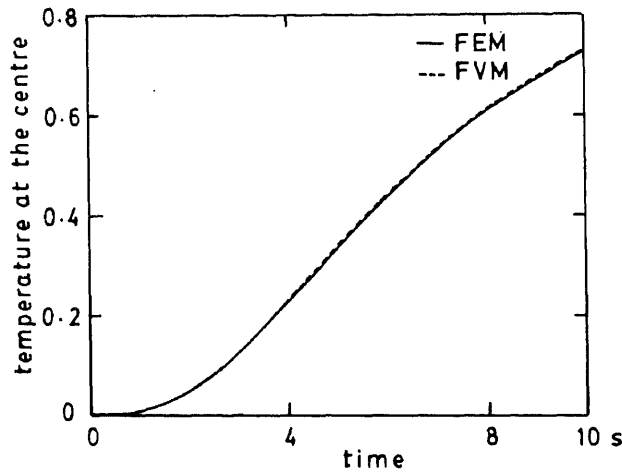


ment model of one quarter plate

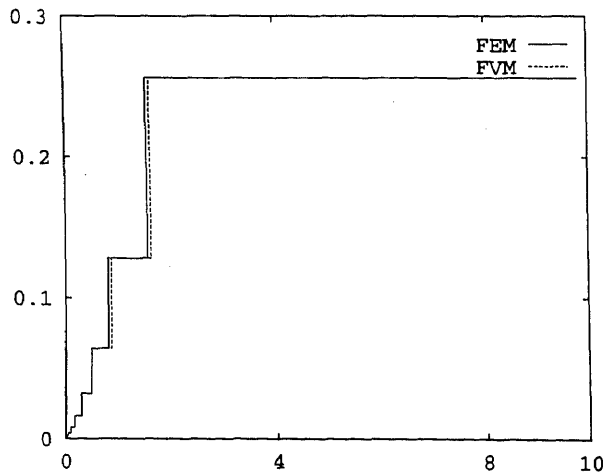
**Figure 9.** Two



**Figure 10.** Nonlinear thermal properties. Temperature dependent  $K$  (a),  $h$  (b) and  $\varepsilon$  (c).



**Figure 11.** Comparative transient temperature of the centre.



**Figure 12.** Adaptive time step variations.

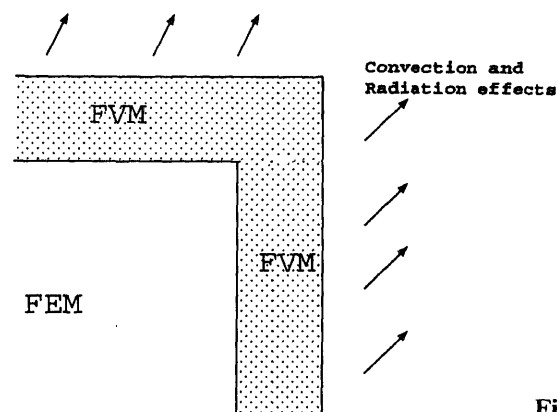
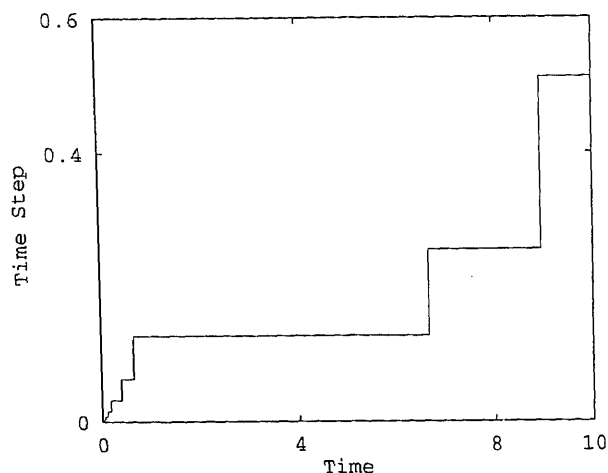


Figure 13. Partitioned meshes.

### Partitionings: Finite element and finite volume partitions

kin finite element method based on the differential form of the global ion of energy provides numerical approximations which are in general conservative. Such numerical approximations however may violate local ive properties, which the finite volume type elements preserve, as these have oped from the local conservation form of the energy equation. Considering al situations, with Galerkin finite elements being employed, local conserva- re likely to be violated in regions where there are external heat fluxes acting rface of a body. Such cases may lead to solutions which may violate some principles (Banaszek 1984). In these instances, the locally conservative finite element is cited (Banaszek 1984) to avoid such violations while preserving energy conservation. Hence, it is proposed that there be mesh partitionings the physical situation to employ locally conservative finite volume elements gions where external heat fluxes are present, and to use the Galerkin finite over the other regions of the domain. The example problem in which the s are subjected to heat fluxes due to convection and radiation is modelled these element partitionings. Finite volume elements are used in the regions he sides which are subjected to convection and radiation heat fluxes and ining region is modelled with Galerkin finite elements. Such a mesh is n figure 13, and such a partitioning approach based on two different ogies will lead to better preservation of local conservation near the external : regions while satisfying global conservation. A time step of 0.05 seconds was employed in the analysis. The adaptive time step variations for this case n in figure 14.

on the proposed hypothesis, improved numerical approximations for general sfer situations can be obtained by employing mesh partitionings based on lly conservative finite volume elements in the regions of external fluxes and gradient fluxes to avoid violation of physical laws and preserve local conserva- llerkin finite elements are employed in other regions. Such modelling and -making depends on the intuition, experience and knowledge of the analyst, ntly efforts are underway to establish more rigour for such partitioning



**Figure 14.** Adaptive time step variations with partitioned meshes.

### Concluding remarks

An adaptive time stepping strategy, based on *a posteriori* error estimates which permits control of global error for finite element/finite volume computations, was described for applicability to general transient thermal problems. To accurately predict the solution behaviour and to effectively make use of computational resources, such approaches are attractive and involve the use of an optimal number of time steps as opposed to that when a single uniform time stepping is used. Mesh partitioning techniques for combining Galerkin finite elements and finite volume based element approaches were proposed to obtain numerically improved physical representations. Test cases showed that the present formulations can accurately track transient behaviour and that the mesh partitionings provide features for improved numerical representations of the physical situations.

The authors are very pleased to acknowledge support, in part, by NASA-Johnson Space Center/Lockheed Engineering and Space Sciences Co., Houston, Texas, and, by the Army High Performance Computing Research Center (AHPCRC), at the University of Minnesota on a contract from the Army Research Office. Additional support and computing times were furnished by Minnesota Supercomputer Institute at the University of Minnesota. Special thanks are due to Mr Tian Ouyang, University of Minnesota.

### References

- Baliga B R, Patankar S V 1983 A control volume finite element method for two dimensional fluid flow and heat transfer. *Numer. Heat Transfer* 6: 263–282
- Banaszek J 1984 A conservative finite element method for heat conduction problems. *Int. J. Numer. Methods Eng.* 20: 2033–2050
- Bathe K J 1982 *Finite element procedures in engineering analysis* (Englewood Cliffs, NJ: Prentice-Hall)
- Eriksson K, Johnson C 1987 Error estimators and automatic time step control for nonlinear



problems-I. *SIAM J. Numer. Anal.* 24: 12-23

1987 *The finite element method* (Englewood Cliffs, NJ: Prentice-Hall)

8 Error estimates and adaptive time-step control for a class of one-step methods in ordinary differential equations. *SIAM J. Numer. Anal.* 25: 908-926

Wendroff B 1964 Difference schemes for hyperbolic equations with high order accuracy. *Commun. Pure Appl. Math.* 17: 381

10, Tamma K K 1991 Applicability/evaluation of flux based representations for second order elements for heat transfer in structures: Generalized  $\gamma_T$ -family. In *AIAA Aerospace Sciences Meeting*, Reno, Nevada

11 Tamma K K 1992 On adaptive time stepping approaches for thermal solidification problems. In *Proc. ASME Winter Annual Meeting, Anaheim, CA, ASME AMD*, vol. 157

12 Damjanic F 1983 Reduced numerical integration in thermal transient finite element analysis. *Comput. Struc.* 17: 261-276

13 1980 *Numerical heat transfer and fluid flow* (New York: McGraw-Hill)

14, Baliga B R 1978 A new finite-difference scheme for parabolic differential equations. *Heat Transfer* 1: 27-37

15 D, Morton K W 1967 *Difference methods for initial value problems*. Interscience Publishers and Applied Mathematics-Number 4 (New York: Interscience)

16 1982 Control volume based finite element formulation of the heat conduction problem. In *3rd AIAA/ASME Joint Thermophysics, Fluids, Plasma and Heat Transfer Conference*, St. Louis, MO

17, Zedan M 1983 Control volume based finite element formulation of the heat conduction equation. *Prog. Astronaut. Aeronaut.* 86: 305-327

18 Namburu 1989 Explicit Lax-Wendroff/Taylor-Galerkin second-order accurate schemes involving flux representations for effective finite element thermal analysis. In *1st Aerospace Sciences Meeting*, Reno, Nevada, AIAA-89-0521

19 Iienkiewicz O C, Baynham J M 1983 Mixed and irreducible formulations in finite element analysis. *Hybrid and mixed finite element methods*. (New York: John Wiley and Sons)

20 Hanson C, Nie Y 1990 An a posteriori error estimate and adaptive time step control for backward Euler discretization of a parabolic problem. *SIAM J. Numer. Anal.* 27: 107-120

21, Curry D M 1977 An implicit-iterative solution of the heat conduction equation with boundary condition. *Int. J. Numer. Methods Eng.* 11: 1605-1619

22 O C, Cheung Y K 1965 Finite elements in the solution of field problems. *The Engineer* 507-510



## Finite element analysis of internal flows with heat transfer

M SRINIVAS<sup>1</sup>, M S RAVISANKAR<sup>1+</sup>, K N SEETHARAMU<sup>1</sup> and  
P A ASWATHANARAYANA<sup>2\*</sup>

<sup>1</sup>Department of Mechanical Engineering, and

<sup>2</sup>Department of Applied Mechanics, Indian Institute of Technology,  
Madras 600 036, India

<sup>+</sup> Present address: Department of Mechanical Engineering, Texas A&M  
University, USA

**Abstract.** This paper presents a finite element-based model for the prediction of 2-D and 3-D internal flow problems. The Eulerian velocity correction method is used which can render a fast finite element code comparable with the finite difference methods. Nine different models for turbulent flows are incorporated in the code. A modified wall function approach for solving the energy equation with high Reynolds number models is presented for the first time. This is an extension of the wall function approach of Benim and Zinser and the method is insensitive to initial approximation. The performance of the nine turbulent models is evaluated by solving flow through pipes. The code is used to predict various internal flows such as flow in the diffuser and flow in a ribbed channel. The same Eulerian velocity correction method is extended to predict the 3-D laminar flows in various ducts. The steady state results have been compared with benchmark solutions and the agreement appears to be good.

**Keywords.** Finite element method; internal flows; Eulerian velocity correction method; diffuser flow; inlet velocity distortion; heat transfer augmentation; 3-D laminar flows; flow through ducts; turbulence models.

### 1. Introduction

Internal flows, such as the flow through straight and curved ducts, cascades, diffusers, nozzles, combustion chambers, turbomachinery stators and rotors are among the most complex flows encountered in practice. The fluid dynamics and heat transfer behaviour of laminar and turbulent flows in channels and ducts are of interest because of wide applications in heat exchangers. The analysis of the hydrodynamics and heat transfer for flow in non-circular ducts is generally more complicated than in the case of circular pipe flow. For example the determination of developing flow, prediction of local and fully developed friction factors and Nusselt number requires a three-dimensional analysis.

Most flows of technological interest are turbulent, at least in some regions. For

many of these flows, relatively simple prediction methods suffice to produce results of engineering accuracy. Turbulence is one of the least understood phenomenon in fluid dynamics. The boundary layer in the earth's atmosphere, water currents below the surface of oceans, flow of water in rivers and canals, flow of fluids in heat exchangers and most combustion processes are some examples of turbulent flows.

Heat transfer augmentation studies in internal geometries with obstructions are of great relevance to the design of heat exchangers. For example, a two-dimensional flow over parallel ribs mounted on the surface of tubes or plates gives rise to separation and recirculation of the flow, which promote turbulent mixing and thereby enhance the rate of heat transfer. These advantages may be offset by an increase in pressure drop, with consequent increase in pumping power. The ability to predict these flows would, therefore, assist in optimizing the design of these devices.

Interest in heat exchanger surfaces with a high ratio of heat transfer area to core volume, as in the case of compact heat exchangers, is increasing at an accelerated pace. Because of the smaller flow passage hydraulic radius the heat exchanger design range usually falls well within laminar flow regime. A common understanding is that turbulent flows provide high heat transfer coefficients and hence are desirable in heat exchanger applications. Laminar flow heat exchangers can also offer substantial weight, volume, space and cost savings. Hence the knowledge of fluid friction and heat transfer in ducts of various flow cross-section geometries is important. In addition to compact heat exchangers, applications of laminar flow theory are also of interest in the aerospace, nuclear, biomedical, electronics and instrumentation fields.

Finite element method (FEM) is capable of handling complex geometries with ease and it is versatile in dealing with the mixed boundary conditions. FEM has an enviable generality in its approach. However, it is not as fast as finite difference techniques due to its lengthy assembly procedure. Attention is focussed on an algorithm which can render a fast finite element code in this work. Heat transfer augmentation by introducing obstructions in the flow path can be studied conveniently using FEM. Thus different types of obstructions with complicated shapes can be easily investigated.

The present work is aimed at predicting internal two- and three-dimensional flows. General purpose finite element codes for 2-D and 3-D have been developed. Nine different turbulence models have been incorporated in the code for the prediction of turbulent flows in 2-D.

The 2-D code is developed for axisymmetric geometries. It can also be used to solve plane problems with a simple idea. A straight line is nothing but part of a circle whose radius tends to infinity. A plane problem can also be viewed as an axisymmetric problem whose axis of symmetry lies far away, in principle, at infinite distance. Numerically infinity can be specified only as a big number. Hence by shifting the value of radial coordinate by a large number, the accuracy of calculation will not be affected. This approach is a valid short cut for solving a plane problem with a code for axisymmetric geometry without altering the code. The 3-D laminar code is developed in Cartesian coordinates.

## 2. Survey of previous work

### 2.1 Turbulent flows

Turbulent flows are characterized by their randomness. The diffusivity of turbulence, which causes rapid local mixing and increased rates of momentum, heat and mass

transfer, is another important feature of turbulent flows. Although the Navier-Stokes equations have been assumed to apply in principle, equally to laminar and turbulent flows, the character of the small-scale details of turbulence prevents a complete analysis when using the present day computers. In order to achieve approximate solutions turbulent flow is often described in terms of averaged quantities. The process of averaging, however, necessitates the appearance of fluctuating velocities in conservative equations. No direct way of knowing the magnitudes of these terms is currently available. This leads to the well-known turbulence closure problem. To overcome this difficulty various turbulence models have been proposed.

A good number of review articles on modelling and computer simulation of turbulent flows is available in the literature (Mellor & Herring 1973; Reynolds 1978, pp. 193-231; Bradshaw *et al* 1981; Haines 1982; Johnson & Launder 1982; Launder 1982; Rodi 1982, 1984; Lumley 1983; Ferziger 1987; Hussaini & Zang 1987; Hutton & Smith 1987, pp. 289-310; Nallasamy 1987; Deissler 1988, pp. 1153-81; Markatos 1988, pp. 1221-75; Murphy 1988, pp. 1131-51; Lakshminarayana *et al* 1989). Nallasamy (1987) gives an excellent review of turbulent flows in internal geometries. Finite element simulation of turbulent flow is discussed at length by Hutton & Smith (1987, pp. 289-310).

One-point turbulence closure models are the most popular ones among all levels of turbulence modelling. These models are based on the averaging of Navier-Stokes equations. There are different levels of complexity in one-point closure, ranging from mixing length models to stress equation models. Many review papers give an extensive introduction to the models coming under one-point closure (Mellor & Herring 1973; Rodi 1982, 1984; Ferziger 1987; Nallasamy 1987; Markatos 1988, pp. 1221-75; Murphy 1988, pp. 1131-51).

All the one-point closures are valid only in the fully turbulent regions. These models neglect the effect of molecular viscosity in comparison with the eddy viscosity. Such an assumption is valid only in fully turbulent regions. There are two distinct approaches to overcome this drawback. The wall function method is also called the high Re version models, while the other one, incorporating damping functions in the model, is also called the low Re version of the turbulence models. In the present code both high and low Re versions of  $k-1$ ,  $k-\epsilon$  and  $q-f$  models have been incorporated.

## 2.2 Turbulent heat transfer modelling

The current status of turbulent heat transfer is reviewed well by Launder (1984, 1988). Pletcher (1988) focusses the attention of his review on forced convection heat transfer. The reviews by Michelic & Wingerath (1988, pp. 1393-1428), Patankar (1988) and Shih (1985, 1987, 1989) give an overall picture of the heat transfer research. The present work confines itself to forced convection heat transfer, where heat is treated as a passive scalar. The wall function methods used to specify the velocity at the wall layer (Benim & Zinser 1985), for the high Re version of the models are utilised for the specification of temperature also for the first time. A wall heat flux temperature  $T_*$  is defined on similar lines to  $U_*$ .

## 2.3 3-Dimensional laminar flows

Although a large number of numerical methods have been proposed with the progress in computers, the transient 3-D analysis of flow remains a difficult task due to the

enormous amount of computational storage and effort required. The explicit time integration scheme has the advantage of saving computational storage owing to mass lumping. The implicit time integration scheme is reported to be more stable than the explicit scheme but the inversion and storage of large matrices are required.

With the advent of modern electronic computers, several different finite-difference formulations of the steady-state, three-dimensional Navier–Stokes equations have been investigated. Difficulties in imposition of the pressure boundary conditions and satisfaction of the continuity equation are encountered in the primitive variables approach. General numerical marching procedures for the calculation of the transport processes in three-dimensional ducts have been proposed by Caretto *et al* (1972) and Curr *et al* (1972).

Gresho *et al* (1981) developed a finite element program for the time-dependent solution of the 3-D Navier–Stokes equations. The model is based on the Galerkin approximation of the primitive variable formulation of the Navier–Stokes equations. Reddy (1982) presented results of a penalty finite element analysis for three-dimensional incompressible flows in enclosures.

In the present study, a general purpose finite element code for full Navier–Stokes equation to predict the recirculating flows is developed. The explicit time-integration scheme is used to save computer storage and computational time.

### 3. Fractional step methods

There are a variety of fractional step methods which can be obtained by appropriately combining the pressure, viscous and the convective terms in the equations. Ferziger (1987) gives an excellent review on the different numerical schemes available. Ravikumaur (1988) uses a fully explicit scheme with mass lumping of the matrices. Mass lumping of the matrices and explicit scheme render a well matched technique (Donea *et al* 1982). Mass lumping shifts the frequency downwards and the explicit scheme upwards, thereby each balancing the other. Reduced order integration is used for pressure (Smith 1984; Benim & Zinser 1985; Autret *et al* 1987; Gresho & Sani 1987; Utnes 1988). Ravikumaur (1988) uses equal-order interpolation with the Eulerian velocity correction scheme. Chequer-board splitting (or spurious pressure oscillations) is a ubiquitous phenomenon. Averaging the pressure field is often resorted to for presenting the results. Specification of the pressure boundary condition is well discussed in Gresho & Sani (1987). Traction-free exit boundary condition is often preferred.

#### 3.1 Governing equations

The non-dimensionalised governing equations for unsteady-state incompressible flow are given below. The index  $j$  is 3 for 3-D equations and 2 for 2-D problems. For 2-D plane problems  $X_2$  becomes very large and its contribution is thus negligible. Also the equations should be read without  $X_2$  for 3-D problems.

*Continuity equation*

$$\frac{1}{X_2} \left[ \frac{\partial (X_2 U_i)}{\partial X_i} \right] = 0. \quad (1)$$

Momentum equations (In  $X_i$  direction)

$$\frac{\partial U_i}{\partial t} + U_j \frac{\partial U_i}{\partial X_j} = \frac{\partial P}{\partial X_i} + \frac{1}{X_2} \frac{\partial}{\partial X_j} \left[ X_2 \left( \frac{1 + \nu_T}{\text{Re}} \right) \left( \frac{\partial U_i}{\partial X_j} + \frac{\partial U_j}{\partial X_i} \right) \right] - \delta_{i,2} \frac{2U_i}{X_2^2} \left( \frac{1 + \nu_T}{\text{Re}} \right), \quad (2)$$

for  $j = 1, 2, 3$ ,

where  $P$  is pressure and  $\nu_T$  is turbulent momentum diffusivity. This is zero for laminar flow problems. Appropriate reference values of velocity  $U_{\text{ref}}$  and length  $L_{\text{ref}}$  along with the kinematic viscosity  $\nu$ , density  $\rho$  and thermal diffusivity  $\alpha$  are used so that the variables are dimensionless. Time is non-dimensionalised by  $L_{\text{ref}}/U_{\text{ref}}$ .

Reynolds number is defined by

$$\text{Re} = U_{\text{ref}} \times L_{\text{ref}}/\nu. \quad (3)$$

Energy equation

$$\frac{\partial T}{\partial t} + U_i \frac{\partial T}{\partial X_i} = \frac{1}{X_2} \frac{\partial}{\partial X_i} \left[ X_2 \left( \frac{1 + \alpha_T}{\text{Pe}} \right) \left( \frac{\partial T}{\partial X_i} \right) \right], \quad (4)$$

where  $\alpha_T$  is turbulent thermal diffusivity which is zero for laminar flow problems, and  $\text{Pe}$  is Peclet number.

Even though nine models are used to study the flow through the pipe, only the details of  $k$ - $\varepsilon$  model are given which is widely reported in literature. Lam & Bremhorst (1981) damping is used for low  $\text{Re}$  version of  $k$ - $\varepsilon$  model.

**$k$ - $\varepsilon$  model:** The turbulent momentum diffusivity is obtained through the solution of two partial differential equations. Lam & Bremhorst's (1981) model is used.  $k$  is non-dimensionalised by  $U_{\text{ref}}^2$  and  $\varepsilon$  by  $U_{\text{ref}}^3/L_{\text{ref}}$ .

**$k$  equation:**

$$\frac{\partial k}{\partial t} + U_i \frac{\partial k}{\partial X_i} - \frac{1}{X_2} \frac{\partial}{\partial X_j} \left[ X_2 \frac{1}{\text{Re}} \left( 1 + \frac{\nu_T}{\sigma_k} \right) \left( \frac{\partial k}{\partial X_j} \right) \right] - Q_k = 0, \quad (5)$$

where  $Q_k$  is the source term given by

$$Q_k = \nu_T \frac{\partial U_i}{\partial X_j} \left[ \frac{\partial U_i}{\partial X_j} + \frac{\partial U_j}{\partial X_i} \right] - \varepsilon. \quad (6)$$

**$\varepsilon$  equation:**

$$\frac{\partial \varepsilon}{\partial t} + U_i \frac{\partial \varepsilon}{\partial X_i} - \frac{1}{X_2} \frac{\partial}{\partial X_j} \left[ X_2 \frac{1}{\text{Re}} \left( 1 + \frac{\nu_T}{\sigma_\varepsilon} \right) \left( \frac{\partial \varepsilon}{\partial X_j} \right) \right] - Q_\varepsilon = 0, \quad (7)$$

where  $Q_\varepsilon$  is the source term given by

$$Q_\varepsilon = C_\mu C_{\varepsilon 1} f_\mu f_1 k \frac{\partial U_i}{\partial X_j} \left[ \frac{\partial U_i}{\partial X_j} + \frac{\partial U_j}{\partial X_i} \right] - C_{\varepsilon 2} f_2 \varepsilon^2/k, \quad (8)$$

$C_\mu, C_{\varepsilon 1}, C_{\varepsilon 2}, \sigma_k$  and  $\sigma_\varepsilon$  are constants equal to 0.09, 1.44, 1.92, 1.0 and 1.3 respectively. The damping functions  $f_\mu, f_1$  and  $f_2$  are given by

$$f_\mu = (1 - \exp(-0.0165 \text{Re}_k))^2 (1 + 20.5/\text{Re}_T), \quad (9)$$

$$f_1 = 1 + (0.05/f_\mu)^3, \quad (10)$$

$$f_2 = 1 - \exp(-\text{Re}_T^2). \quad (11)$$

$\text{Re}_T$  and  $\text{Re}_k$  are defined by the following equations

$$\text{Re}_T = \text{Re } k^2/\varepsilon, \quad (12)$$

$$\text{Re}_k = \text{Re } k^{0.5} \cdot y. \quad (13)$$

Turbulent momentum diffusivity is then calculated by

$$\nu_T = \text{Re } f_\mu C_\mu k^2/\varepsilon. \quad (14)$$

### 3.2 Method of solution

The Eulerian velocity correction method, a solution algorithm for unsteady, incompressible Navier–Stokes equations, is used to solve the momentum equations. Solution is advanced in three steps within every time step. The three steps involved are calculation of pseudo velocities, calculation of pressure from the Poisson equation and correction of pseudo-velocities to obtain velocities at the next time step.

*Step 1. Calculation of pseudo velocities:* The pseudo velocities  $V_i$  are calculated from (2) by dropping the pressure terms. Since these velocities will not satisfy the continuity equation, they are known as pseudo velocities.

$$\begin{aligned} \frac{\partial V_i}{\partial t} = & -U_j \frac{\partial U_i}{\partial X_j} + \frac{1}{X_2} \frac{\partial}{\partial X_j} \left[ X_2 \left( \frac{1 + \nu_T}{\text{Re}} \right) \left( \frac{\partial U_i}{\partial X_j} + \frac{\partial U_j}{\partial X_i} \right) \right] \\ & - \delta_{i,2} \frac{2U_i}{X_2^2} \left( \frac{1 + \nu_T}{\text{Re}} \right), \quad \text{for } j = 1, 2, 3. \end{aligned} \quad (15)$$

An explicit Euler's scheme is used to expand the time derivative of the above equation in the time domain,

$$\frac{\partial V_i}{\partial t} = \frac{V_i^{n+1} - U_i^n}{\Delta t^n}. \quad (16)$$

From (15) and (16) pseudo velocity  $V_i$  can be calculated.

*Step 2. Pressure Poisson equation:* By modifying (2) and (15), the following equation is obtained,

$$\frac{\partial U_i}{\partial t} - \frac{\partial V_i}{\partial t} = - \frac{\partial P}{\partial X_i}. \quad (17)$$



Both the time derivatives are expanded explicitly,

$$\frac{\partial P^{n+1}}{\partial X_i} = \frac{V_i^{n+1} - U_i^{n+1}}{\Delta t^n}. \quad (18)$$

Taking the partial derivative of (18) with respect to  $X_i$  and after some modification, the following pressure Poisson equation is obtained,

$$\frac{1}{X_2} \left[ X_2 \frac{\partial^2 P^{n+1}}{\partial^2 X_i} \right] = \frac{1}{X_2} \frac{1}{\Delta t^n} \left( \frac{\partial(X_2 V_i^{n+1})}{\partial X_i} \right). \quad (19)$$

**Step 3. Velocity correction:** The original velocities of the next time step  $U_i^{n+1}$  are obtained by correcting the pseudo velocities using the evaluated pressure field

$$U_i^{n+1} = V_i^{n+1} - \Delta t^n (\partial P^{n+1} / \partial X_i). \quad (20)$$

Some of the advantages of the present scheme are – the pressure Poisson equation alone is solved from a set of algebraic equations. In steps 1 and 3 mass lumping is done which makes the stiffness matrix diagonal. The inverse of the diagonal matrix is just the inverse of each element which saves the computational time. The explicit scheme tends to shift the frequency of oscillation up and the mass lumping procedure shifts it down. Hence the combination of these two will result in a well-matched scheme. The stiffness matrix in the Poisson equation does not depend on anything that evolves with time and hence the assembly procedure is done only once in the first iteration.

The solution of the partial differential equation is sought using the finite element method. The Galerkin weighted-residual technique is used to formulate the problem. Benim & Zinser (1985) have reported that linear elements are preferable to higher order elements for turbulent flow problems. Based on their suggestion, linear triangular elements in two-dimensional problems and tetrahedron elements for three-dimensional problems are used. Also triangular elements and tetrahedral elements do not require any numerical integration which again saves computational time.

## 4. Results and discussion

### 4.1 2-D turbulent flows

**4.1a Developing flow through a smooth circular pipe:** A circular pipe represents the simplest axisymmetric geometry. Turbulence modelling is not complete even in this geometry. Comparative study of different models of turbulence by the same code, for a particular problem under identical conditions, would throw unambiguous light on the strength and weakness of each model. In the present section, different models of turbulence are evaluated for their predictive capabilities. Both momentum and heat transfer are studied.

Taylor *et al* (1977) suggest the use of special elements with logarithmic shape functions near the wall to get better predictions. High Re versions of the model tend to give accurate results for this problem (Benim & Zinser 1985; Morgan *et al* 1987;

Taylor *et al* 1981, pp. 341–9). However, Martinuzzi & Pollard (1989) conclude that the low Re version  $k$ - $\varepsilon$  model performs better than the high Re version model. Benim & Zinser (1985) use the high Re version of the  $k$ - $\varepsilon$  model for their predictions. One of their important suggestions is the use of linear elements in favour of higher order elements. Taylor *et al* (1977) clearly bring out that in high Re version models, the initial assumption of wall shear stress is very crucial. The predictions are highly sensitive to the initial assumption. Benim & Zinser (1985) suggest a brilliant approach, by which this problem can be completely obviated. In the present work, this approach is extended to heat flow prediction for all the boundary conditions. The conclusion of Martinuzzi & Pollard (1989), that low Re models perform better is not true over a range of Reynolds number, particularly at high Reynolds numbers. The low Re model used by them, which is based on Lam & Bremhorst (1981) damping, is known for its accurate prediction near the wall, especially at low Reynolds numbers (less than  $10^5$ ).

(i) *Geometry and boundary conditions* – The developing length of turbulent flow in a pipe is approximately 30–50 pipe diameters. Hence the length of the pipe is selected as 70 times the pipe diameter.

Inlet values of turbulent kinetic energy  $k$  and its dissipation are specified based on Martinuzzi & Pollard (1989). The turbulent intensity of the stream at the inlet is assumed to be 6%, and a uniform  $k$  profile is specified. The inlet condition of  $\varepsilon$  is calculated from the following relationship.

$$\varepsilon_{Ln} = (C_\mu \cdot k^{3/2}) / (0.015D). \quad (21)$$

(ii) *Results and discussion* – Table 1 compares the fully developed friction factor, Nusselt number and the centreline velocity. All the derivatives are calculated using the wall function approach. The comparison is shown for Reynolds number 100 000. Performance of the low Re version of the  $q$ - $f$  model is bad. In all other cases the error is well within the acceptable limit.

**Table 1.** Comparison of fully developed centre line velocity, friction factor and Nusselt number for flow through a smooth circular pipe at  $Re = 100\,000$  and  $Pr = 0.71$ .

Model	$U_{CL}$		$C_f$		$Nu$	
	Present work	Error <sup>†</sup> (%)	Present work	Error <sup>†</sup> (%)	Present work	Error <sup>*</sup> (%)
PML	1.150	– 5.7	0.0183	1.7	186	2.0
NML	1.216	– 0.3	0.0169	– 6.1	166	– 8.9
VEV	1.242	1.8	0.0168	– 6.7	163	– 10.6
$k$ -1 (LR)	1.200	– 1.6	0.0182	– 1.1	186	2.0
$k$ -1 (HR)	1.183	– 3.0	0.0173	– 4.2	178	– 2.3
$k$ - $\varepsilon$ (LR)	1.190	– 2.5	0.0182	1.1	195	7.0
$k$ - $\varepsilon$ (HR)	1.160	– 4.9	0.0176	– 2.5	180	– 1.3
$q$ - $f$ (LR)	1.120	– 8.2	0.0250	38.9	240	31.7
$q$ - $f$ (HR)	1.160	– 4.9	0.0190	5.6	197	8.1

<sup>†</sup>Compared with the fully developed value 1.220; <sup>‡</sup>Compared with the fully developed value 0.018; <sup>\*</sup>Compared with the fully developed value 182.30

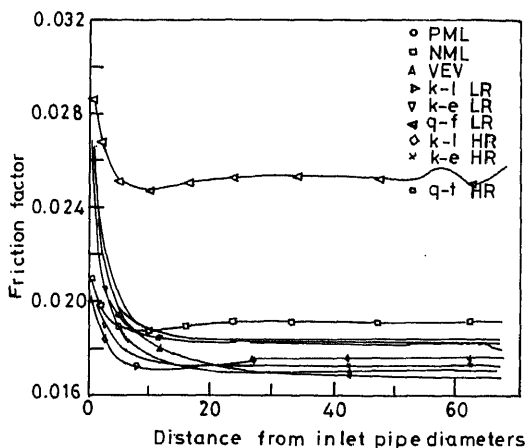


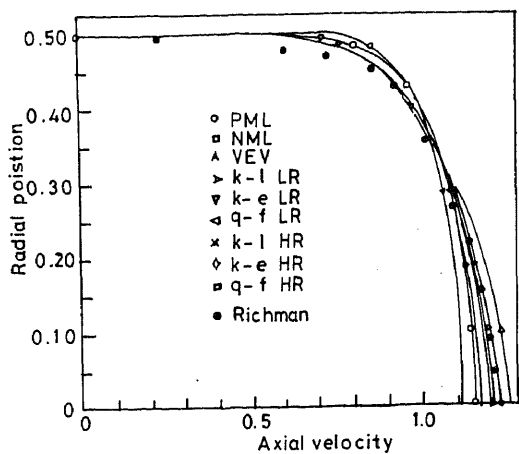
Figure 1. Comparison of friction factor for flow through a pipe at  $Re = 100\,000$ , for different turbulence models.

In the literature, only the high  $Re$  version of the  $q, f$  model is available (Smith 1984). The high  $Re$  version of this model is often quoted for its numerical stability (Hutton & Smith 1987, pp. 289–310). Hence an attempt to develop a low  $Re$  version is made. In the process of derivation a few terms are ignored, these terms are primarily responsible for the bad prediction of the  $q-f$  model. The high  $Re$  version of  $q-f$  model's performance indicates that the ignored terms are not important away from the wall. Tuning the damping constants will improve the results predicted by the  $q-f$  model, as the values are only shifted numerically and the trends agree well with the behaviour of the other models.

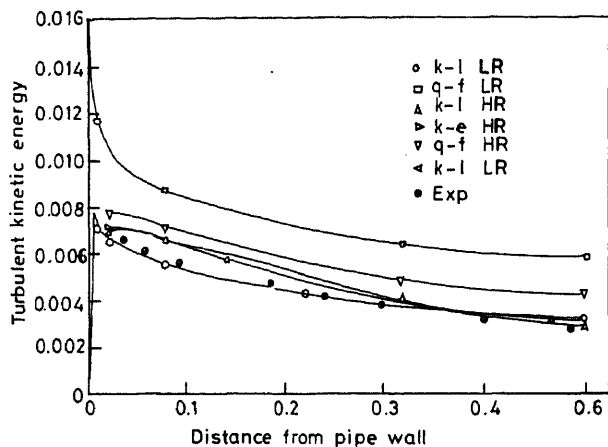
Figure 1 also reveals another interesting feature. But for zero equation models, all the other models show a dip in the friction factor around 10 diameters from the entry. This observation coincides with the results of Benim & Zinser (1985). This undershoot in the curve is not predicted by the zero equation models as they do not incorporate turbulent history in their prediction. As the boundary layer develops along the wall the central core is accelerated. When the flow is accelerated the turbulent kinetic energy decreases. This influences the prediction of the eddy viscosity. A decrease in eddy viscosity in turn decreases the friction factor, which is the cause of the undershoot of friction factor.

Figure 2 compares the fully developed profiles predicted using different models with the experimental results as in Ravisankar (1992). Fully developed turbulent kinetic energy profiles predicted by the high and low  $Re$  versions of  $k-l$ ,  $k-\epsilon$  and  $q-f$  models are compared with the experimental results of Lawn (vide Martinuzzi & Pollard 1989) in figure 3. Low  $Re$  version  $q-f$  model over-predicts turbulent kinetic energy. This is the reason for over-prediction of eddy viscosity, which in turn is responsible for the over-estimation of the skin friction coefficient.

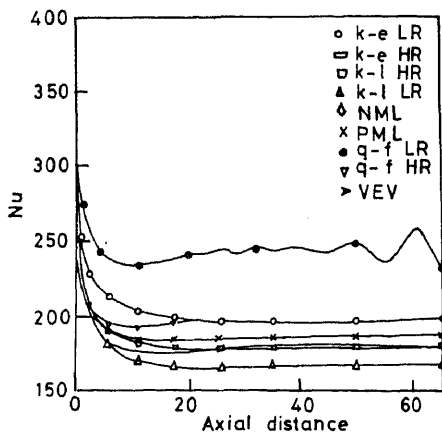
The damping functions used in the models are not valid over a big range of Reynolds numbers. Most of them are valid only for  $Re$  less than 500 000. On the other hand the high  $Re$  version of the models can be applied only in the fully turbulent zones. Figure 4 shows the variation of  $Nu$  along the pipe length, predicted with different models. The trends are similar to the variation of friction factor. Especially the prediction of low  $Re$   $q-f$  model indicates this clearly. Also the undershoot of  $Nu$  is similar to the undershoot of  $C_f$ .



**Figure 2.** Comparison of fully developed velocity profile for flow through a pipe at  $Re = 100,000$ , for different turbulence models.



**Figure 3.** Comparison of fully developed turbulent kinetic energy profile for a flow through a pipe at  $Re = 100,000$ , for different turbulence models.



**Figure 4.** Comparison of Nusselt numbers for flow through a pipe at  $Re = 100,000$ , for different turbulence models.

**Table 2.** Comparison of the number of iterations for convergence using  $k-\epsilon$  model with modified wall function approach for different initial distributions.

S. No.	Wall boundary Condition	Initial Temperature distribution	Iterations <sup>†</sup> required for convergence
1	Temperature	0.0	87
2	Temperature	1.0	91
3	Heat flux	0.0	95
4	Heat flux	1.0	141

<sup>†</sup> Maximum error tolerable between two successive iterations is specified as  $1 \times 10^{-7}$

One of the important contributions of the present work is the extension of the wall function approach of Benim & Zinser (1985) to heat transfer prediction. This reduces the anxiety in specifying the initial temperature distribution at the wall layer. Table 2 gives the details about the number of iterations required for convergence for the correct values of two extreme conditions. A steady-state solver with high Re  $k-\epsilon$  model is used to obtain the solution. The converged values are the same even though the assumptions of initial distributions are drastically different.

(iii) *Conclusions* – (1) The modified wall function approach for heat transfer prediction with high Re version of the models is successful. This approach is insensitive to the initial distribution of wall temperature.

(2) The predictions of low Re version of  $q-f$  model are poor. The terms neglected in the process of derivation are found to be important near the wall.

(3) The one-equation  $k-1$  model, displays excellent stability and the quality of predictions is good. This model is highly reliable for its performance. The only difficulty is the specification of the mixing length distribution for complex geometries.

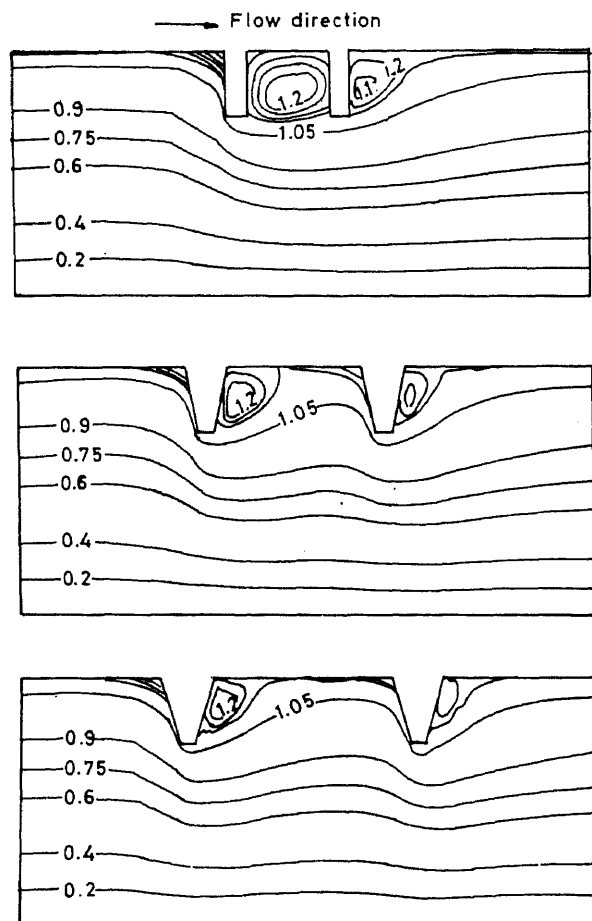
(4) All the damping functions need critical examination. Invariably the low Re versions of the models fail at  $Re = 500\,000$  and above.

**4.1b Heat transfer augmentation in channels – effect of rib wall inclination:** Channels with turbulence promoters such as ribs, fins and thin obstructions are used for heat transfer augmentation in heat exchangers. All the experiments show that the heat transfer coefficient increases by a factor of two in the vicinity of the reattachment point. The flow of cooling air in gas turbine blades can be approximated as the flow through a channel with two opposite rib roughened walls. The wall of the rib is generally made vertical for obtaining maximum mixing. However, slightly inclining the wall of the rib brings down the total pressure drop and results in better heat transfer augmentation efficiency.

Currently, a large number of technical papers are available for turbulent flow inside rib-roughened channels. However little attention is given to the shape of the rib. Han *et al* (1978) experimentally studied the effect of angle of attack, pitch ratio and rib shape on friction factor and heat transfer results. The shape of the rib was altered by filling modelling clay in the corners of the rectangular ribs instead of inclining the wall. The effects of rib shape on the pressure drop and heat transfer augmentation efficiency are not studied.

A brief literature survey is summarized below. Liou *et al* (1990) presented mean velocity and turbulence intensity profiles for an arrangement of two pairs of turbulence promoters mounted in tandem in channel flow. The predictions are compared with experiments. Measured axial velocity and turbulent kinetic energy values are given as inlet conditions for the numerical predictions. The effect of pitch ratio and the influence of  $Re$  on the reattachment length are documented. It has been found experimentally that there are two major enhancement peaks, one located slightly downstream of the leading edge of the rib and the other 0.5 to 1.0 rib heights upstream of the reattachment point. Acharya *et al* (1993) used nonlinear and standard  $k-\epsilon$  models with wall functions for predicting the recirculation lengths and maximum Nusselt number locations and compared them with the experiments. Both models predicted reattachment lengths well but under-predicted local Nusselt numbers. This is attributed to inadequacies in wall functions. Also the Nusselt number peak near the leading edge of the rib is not predicted.

The Reynolds number based on hydraulic diameter of the channel is 85 600. The study is carried out with three pitch ratios (PR), 5, 10 and 15. The rib wall angles ( $\theta$ ) considered are 0, 35 and 50. The aim of the study is to investigate the variations in Nusselt number, friction factor, reattachment lengths and efficiency of heat transfer



**Figure 5.** Stream-line plots for flow through ribbed channels for different pitch ratios and different rib-wall inclinations.

**Table 3.** Comparison of reattachment length for turbulent flow through sudden pipe expansion.

Inlet $k$	Inlet $k$	Inlet $U$ profile	Reattachment length
0.16	0.23	1/7 law	6.10
0.11	0.22	1/7 law	8.60
0.06	0.0882	1/7 law	10.10
0.06	0.1633	1/7 law	5.70
0.06	0.0882	Uniform profile	8.90

augmentation by varying the wall inclination and pitch ratios. The same geometry is used as in Liou *et al* (1990). The reattachment lengths are calculated from the trailing edge of the rib.

(i) *Results and discussion* – Figure 5, shows some typical stream line plots. As the rib wall inclination increases, the separation bubble became smaller.

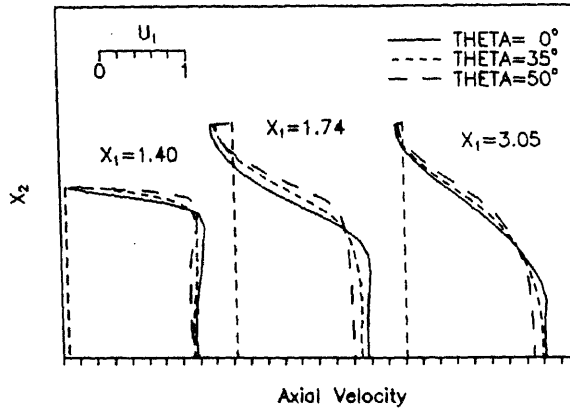
The prediction of reattachment length is very sensitive to the inlet conditions of  $k$  and  $\epsilon$ . For the flow through a sudden pipe expansion the authors studied the effect of inlet conditions of  $k$  and  $\epsilon$  and table 3 gives the prediction of reattachment lengths for  $Re = 100\,000$  with  $k$ - $\epsilon$  model. The experimental value for reattachment length for sudden expansion is around 8.5 to 9.0 step heights (Nallasamy 1987). These results clearly bring out the effect of inlet  $k$  and  $\epsilon$  on the prediction of reattachment length. For the present study no comparison of reattachment length was done since the inlet  $k$  and  $\epsilon$  are not given by Liou *et al* (1990).

Table 4 compares the reattachment lengths for various pitch ratios and step wall inclinations. The reattachment length after second step  $X_{R2}$  is around 3 in all the cases. The rib wall inclination has little effect on the  $X_{R2}$ . Liou *et al* (1990) reported that  $X_{R2}$  remains almost constant for  $5 < PR < 20$ . For  $\theta = 0^\circ$  and PR 10 and 15, the  $X_{R1}$  is almost double that of  $X_{R2}$ .

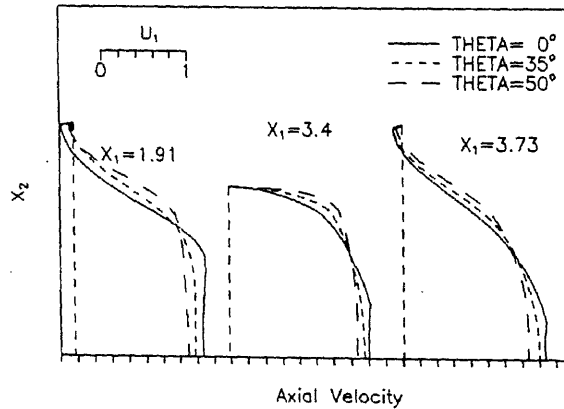
Figures 6 and 7 compare the velocity profiles at three different sections for PR = 10 and 15. As the rib wall inclination increases the negative velocities and centre-line

**Table 4.** Values of  $X_{R1}$  and  $X_{R2}$ .

$\theta$	PR	$X_{R1}$	$X_{R2}$
0	5	—	3.026
	10	5.633	2.972
	15	5.953	3.119
35	5	—	2.689
	10	3.540	2.542
	15	3.659	2.971
50	5	—	2.649
	10	3.611	2.561
	15	3.746	2.756



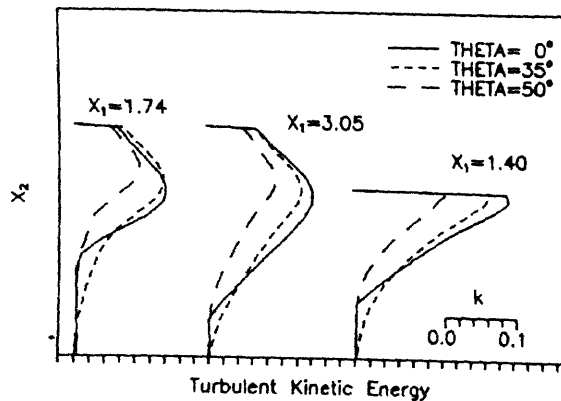
**Figure 6.** Comparison of axial velocity profiles at different locations for different  $\theta$  and  $PR = 10$ , for flow through ribbed channels.



**Figure 7.** Comparison of axial velocity profiles at different locations for different  $\theta$  and  $PR = 15$ , for flow through ribbed channels.

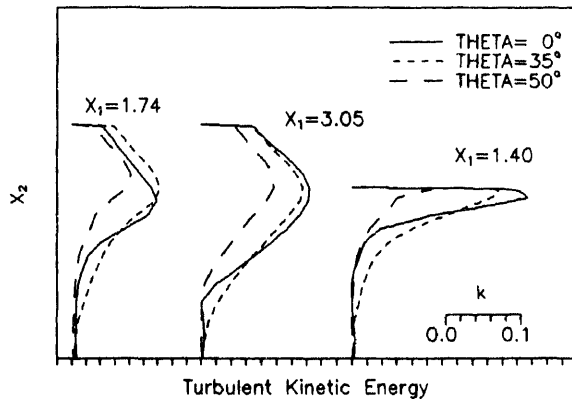
velocities decrease. This explains the reduction in reattachment lengths. Similar flow distributions are found in all cases, downstream of the second rib. Further downstream, the flow eventually approaches that of the fully developed channel flow.

Figures 8 and 9 show the variations of turbulent kinetic energy profiles for  $PR = 10$  and 15. The value of turbulent kinetic energy in the centre region of the flow is almost zero. On the other hand, large values of  $k$  are found in the flow separation region



**Figure 8.** Comparison of turbulent kinetic energy profiles at different locations for different  $\theta$  and  $PR = 10$ , for flow through ribbed channels.

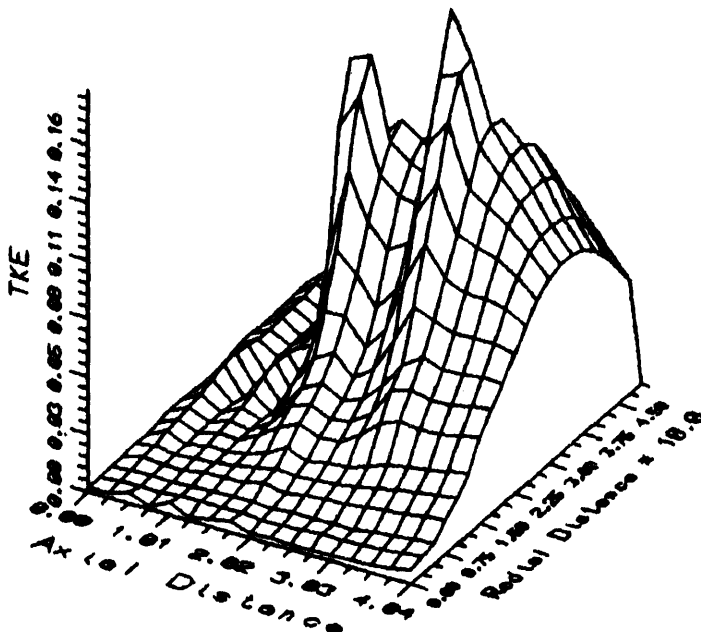




**Figure 9.** Comparison of turbulent kinetic energy profiles at different locations for different  $\theta$  and  $PR = 15$ , for flow through ribbed channels.

and the shear layers where high velocity gradients exist. As the angle  $\theta$  increases, the peak value of turbulent kinetic energy reduces. This is due to the reduced mixing of the flow. Figure 10 shows the three-dimensional plot of turbulent kinetic energy for  $PR = 10$  and  $\theta = 35^\circ$ . The two peaks in the plot correspond to the peak values of kinetic energy below two ribs.

Figures 11 and 12 show the Nusselt number peaks clearly. As was reported by Liou *et al* (1990) every step has two peaks, one just near the leading edge of the step and another 0.5 to 1.0 times the step height upstream of the reattachment point. It can be seen from these figures that both the peaks are predicted well in these cases. However, the peaks occurred at a distance of about 1 to 1.5 times the step height, upstream of the reattachment points.



**Figure 10.** 3-D plot of turbulent kinetic energy for  $PR = 10$  and  $\theta = 10^\circ$ .

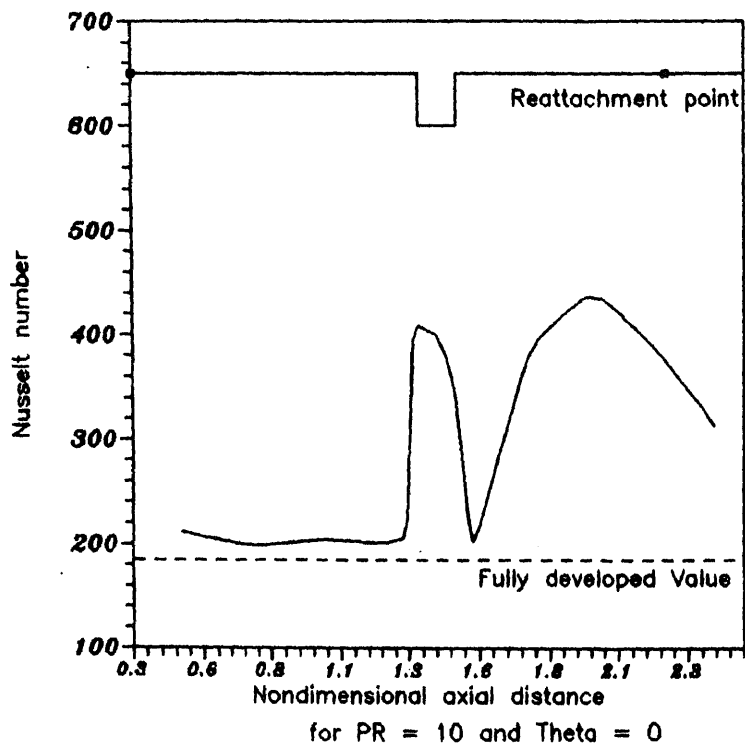


Figure 11. Nusselt number peaks observed for  $PR = 10$ .

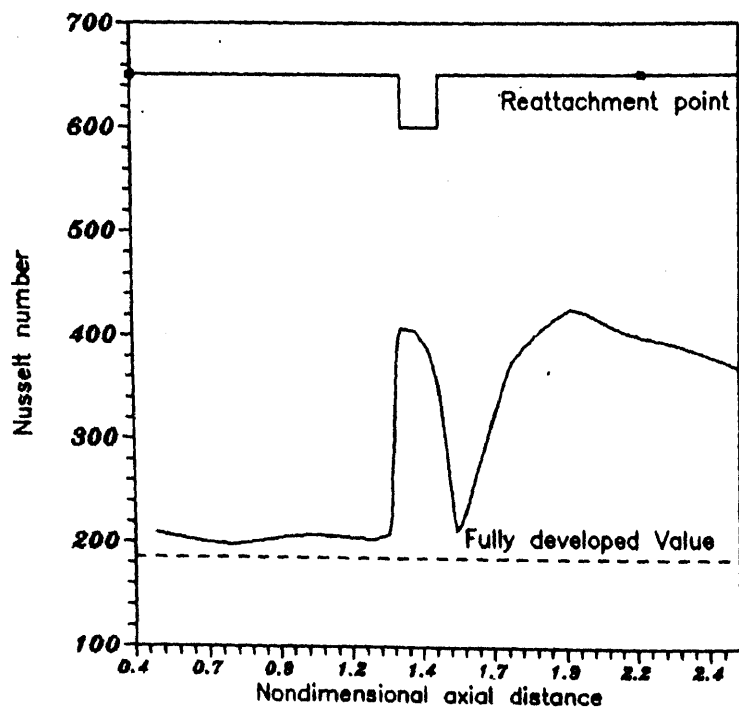


Figure 12. Nusselt number peaks observed for  $PR = 15$ .

The conventional way of estimating the heat transfer augmentation is given by Rabas (1989) as

$$\eta = \frac{\text{ratio of average heat transfer coefficient}}{\text{ratio of pressure drop per unit length}} \quad (22)$$

Geometry without turbulence promoters is taken as the reference for comparison. Table 5 gives the efficiencies for different cases for Pr number = 0.722. Comparisons are made with smooth channel flow keeping all other conditions such as length, Re and inlet profiles the same. The results for smooth channel are also obtained by the same code using the  $k-\epsilon$  model.

Then  $\eta$  reduces with pitch ratio and increases with rib wall angle inclination. The average Nusselt number is reduced with pitch ratio and rib wall angle inclination. The average Nusselt number varies between 1.75 and 2.0 times the smooth channel Nusselt number. Liou *et al* (1990) reported that relative to a smooth duct, the presence of periodic ribs at two opposite walls yields up to 2.2-fold increase in fully developed average Nusselt number.

(ii) *Conclusions* – The pressure drop in channels with ribs reduced with rib wall inclination. There is no significant drop in average heat transfer coefficient. All the trends in local Nusselt number and reattachment lengths are predicted. The improvement in heat transfer augmentation varies between 30 and 39 percent for various step-side inclinations when compared with the straight ribs, the maximum being for PR = 10 and  $\theta = 35^\circ$ . It is also observed that the maximum average Nusselt number is achieved for a pitch ratio of 10.

**4.1c Flow through a plane diffuser – Comparison between flow with and without inlet velocity distortions:** Diffusers are widely used for converting kinetic energy to pressure energy. A proper understanding of flows in a diffuser will lead to a better design of fluid machines with improved efficiency. The flow in a diffuser is highly complex and simple analytical solutions are not available to treat such turbulent flows under adverse pressure gradients. Most of the numerical predictions of diffuser flows are based on boundary layer approximations and solutions are obtained by either solving the integral equations or by solving the modelled differential equations.

It is well known that certain types of inlet velocity distortions affect the efficiency of the diffuser in converting kinetic energy to pressure energy. Hence the influence

**Table 5.** Comparison of  $\eta$  values.

$\theta$	PR	$\overline{Nu}/Nu_s$	$\Delta P/\Delta P_s$	$\eta$
0	5	1.894	9.28	0.2041
	10	1.851	11.41	0.1622
	15	1.795	12.40	0.1448
35	5	1.952	7.65	0.2552
	10	1.918	9.76	0.1965
	15	1.867	9.92	0.1882
50	5	1.737	6.47	0.2685
	10	1.845	8.17	0.2258
	15	1.770	9.11	0.1943

of inlet velocity distortions on diffuser performance needs detailed investigation. It has been observed that with certain types of velocity distortions at the inlet of the diffuser, the efficiency decreases, while with certain other types of distortions, the efficiency improves when compared to the uniform velocity distribution at the inlet.

The prediction of diffuser flows with uniform inlet velocity is fairly well established. Lai & coworkers report the inadequacy of the wall-function approximations for diffuser flows. Since the high Reynolds number versions of the standard two-equation models are based on the local equilibrium assumption and use some kind of wall function to handle the near wall flow, they are not suitable for diffuser flow calculations. In order to account for the anisotropic behaviour near a wall some kind of a low Reynolds number version of a turbulence model capable of resolving the flow up to the wall should be considered.

Hah (1983), used the finite difference method to predict turbulent flows in planar, conical and annular diffusers with inlet swirl and inlet velocity distortion. With the swirling velocity component, the flow is pressed towards the wall by centrifugal force and the wall boundary layer is less likely to separate even if the diffuser divergent angle is large, and a higher pressure recovery coefficient is observed. Inlet velocity distortion is obtained by putting a thin ring outside the wall boundary layer at the inlet for a diffuser of total expansion angle of  $16^\circ$ . The separation near the wall is suppressed when the inlet flow is altered, resulting in more favourable pressure gradient along the wall and higher diffuser performance is thus obtained. Hoffman (1982) also altered the inlet velocity and reported an improvement in the overall diffuser efficiency.

The present investigation is aimed at studying the improvement in diffuser performance by distorting the inlet velocity profile. Chitambar (1978) conducted experiments for a plane diffuser with a divergent angle of  $4^\circ$ , both with and without inlet distortions. No separation is reported in the diffuser. The present results are compared with the experiments conducted by Chitambar (1978). A low Reynolds number version of the  $k-\epsilon$  model is used to close the momentum equations.

(i) *Results and discussion* – Figures 13 and 14 give the stream lines for Case 1 (flow without inlet velocity distortion) and Case 2 (flow with distortion) respectively. The stream lines in Case 2 are slightly packed together in the first half of the diffuser. This is due to the distortion in inlet velocity profile. In the second half of the diffuser

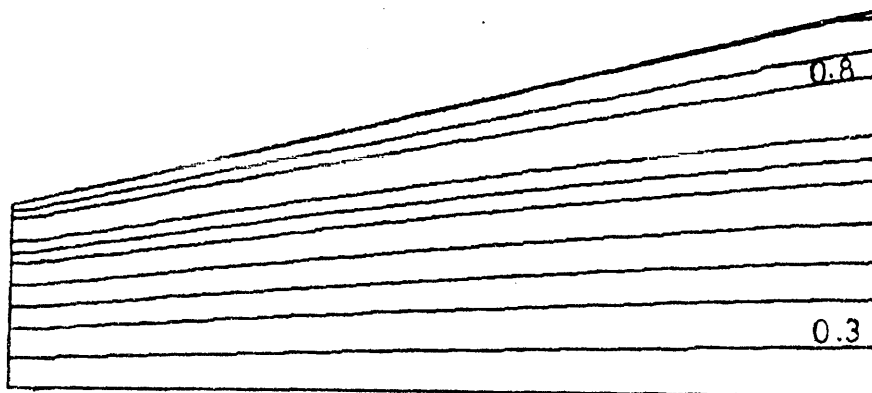


Figure 13. Streamline plot for flow through diffuser without inlet distortion.

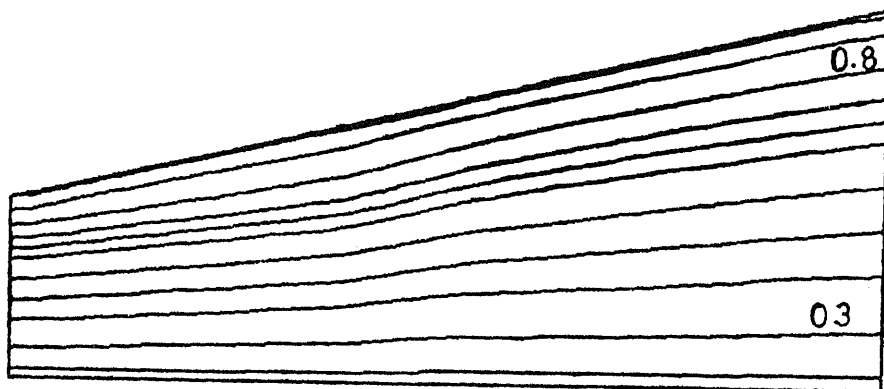


Figure 14. Streamline plot for flow through diffuser with inlet distortion.

the distortion effect dies out and the stream lines become similar to that of the Case 1. No separation is predicted as reported in the experiments. Figures 15 and 16 compare the predicted velocity profiles at three downstream sections with experiments for Case 1 and Case 2. For Case 1 the prediction is good. In the second half of the diffuser the maximum velocity is slightly less than experiments for Case 2. In most of the sections the near-wall prediction is good and the maximum error in the prediction of velocity profiles is 4.1 percent. The distortion in the velocity profile is predicted very well.

The inlet condition for  $\varepsilon$  in Case 1 is calculated using (21) where  $D$  is taken as the width of the diffuser. The inlet condition for Case 2 is calculated using the following formula.

$$\varepsilon_{in} = (C_{\mu}^{3/4} k^{3/2}) / (\kappa y), \quad (23)$$

where  $\kappa$  is a constant in Prandtl mixing length and is equal to 0.4.

The turbulent kinetic energy profiles for Case 1 are given in figure 17. For Case 2 the profiles are compared with experiments in figure 18. In the first half of the diffuser the prediction is good and in the second half the peak value in the profile is slightly under-predicted. The peak values of turbulent kinetic energy are much higher when compared with free stream turbulence. There is marginal increase in the intensity of free stream turbulence, towards the exit of the diffuser. The occurrence of the peak is very close to the wall near the inlet of the diffuser and the peak is shifted away from the wall with distance in the stream-wise direction.

In the calculation of  $\varepsilon_{in}$  from (23), if the normal wall distance is taken from the

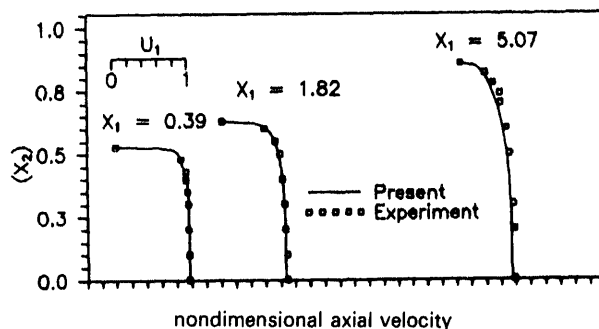


Figure 15. Comparison of axial velocity for flow through diffuser without inlet distortion at different axial locations.

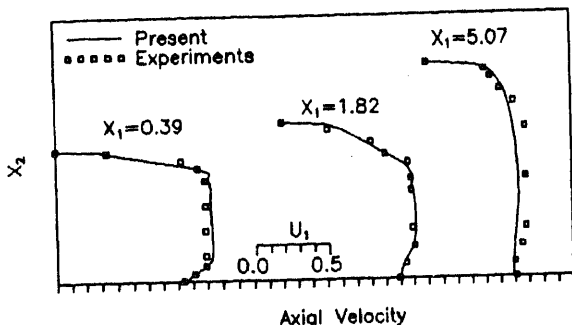


Figure 16. Comparison of axial velocity for flow through diffuser with inlet distortion at different axial locations.

wall of the diffuser the peak value in the  $k$  profile in Case 2 near the axis due to the inlet distortion is under-predicted. This could be due to the fact that the aerofoil also acts as a wall. For the calculation of  $\epsilon_{in}$  the normal wall distance in (23) is slightly modified. The tip of the aerofoil is considered to be a wall and the normal wall distance is taken as the minimum distance from the diffuser wall and the aerofoil tip. This makes the normal distance of nodes near the axis less than that of the distance from the wall and results in higher  $\epsilon$  values near the axis. As a result the peak value in the  $k$  profile has increased and compares well with the experiments in the initial portion of the diffuser.

The prediction of pressure distribution is very vital for the diffuser. Figures 19 and 20 show the comparison of pressure recovery percentage with axial distance. It agrees very well with the experiments. Pressure recovery factor is defined as the ratio of actual pressure recovered to the theoretical pressure recovery possible. The theoretical pressure recovery possible is calculated using one-dimensional Bernoulli's equation between inlet and outlet sections of the diffuser. In Case 2, 81% of the possible kinetic energy has been converted into pressure energy. In Case 2 it is 88%. Figure 21 compares the pressure recovery factor with experiments and the agreement is good. Figure 22 shows the variation of pressure recovery factor for Case 2. It can be seen that 61% of the possible pressure recovery takes place in the first half of the diffuser in Case 1 and 66% in the first half in Case 2. The remaining 39 and 34% recovery takes place in the second half of the diffuser for Case 1 and Case 2 respectively. An improvement of 8.8% in pressure recovery has been obtained in Case 2 from Case 1.

Figure 23 and 24 show the comparison of friction factors for Case 1 and Case 2. The maximum error in this comparison is about 6.82%. For Case 2 the comparison in the first half is better when compared to the second half.

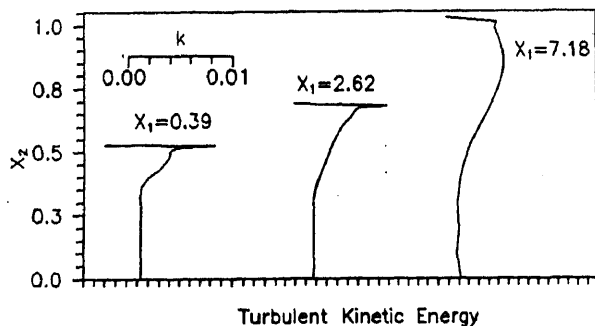


Figure 17. Variation of turbulent kinetic energy for flow through diffuser without inlet distortion at different axial locations.

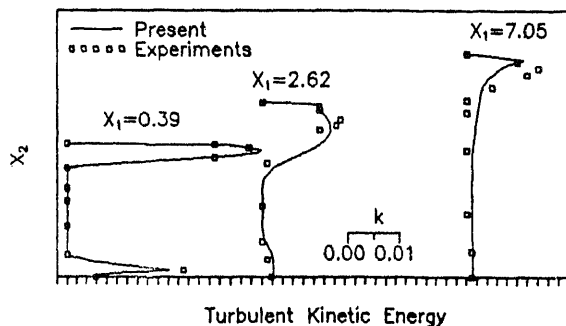


Figure 18. Comparison of turbulent kinetic energy for flow through diffuser with inlet distortion at different axial locations.

(ii) *Conclusions* – The inlet velocity distortion created by the aerofoil resulted in 8.8% more pressure recovery in a two-dimensional plane diffuser with divergent angle  $4^\circ$ . The general purpose finite element code based on low Reynolds number  $k-\epsilon$  equation predicted the mean velocity profiles, kinetic energy profiles and pressure recovery reasonably well and provides useful guidance and information for the advanced design.

#### 4.2 3-D Laminar flows

4.2a *Flows in rectangular ducts:* Numerical solutions of 3-D laminar flows in ducts

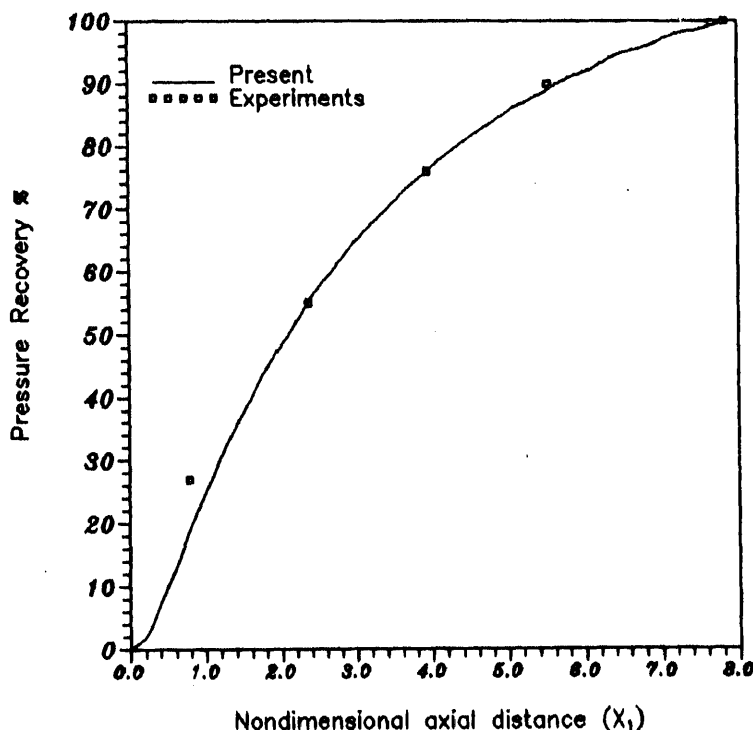


Figure 19. Comparison of pressure recovery percentage for flow through diffuser without inlet velocity distortion.

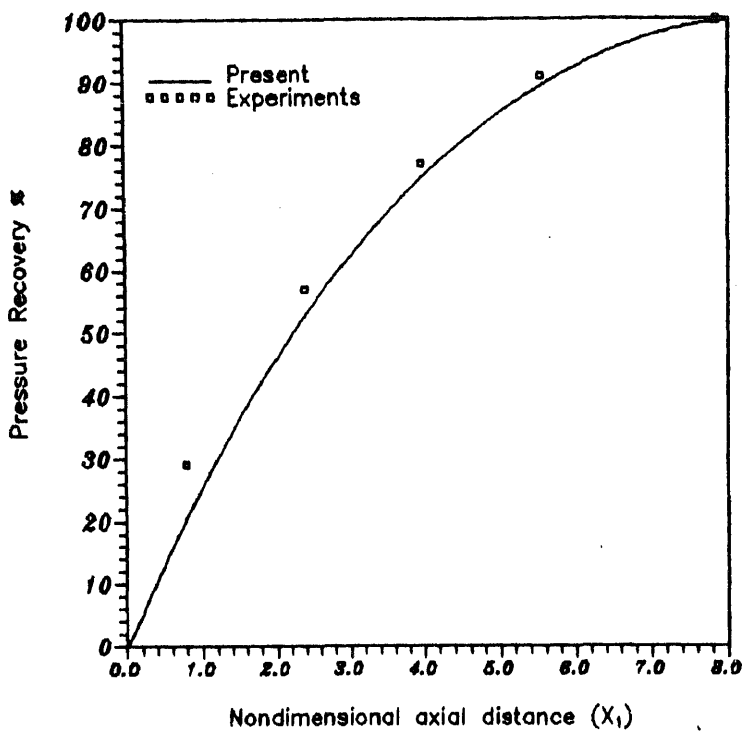


Figure 20. Comparison of pressure recovery percentage for flow through diffuser with inlet velocity distortion.

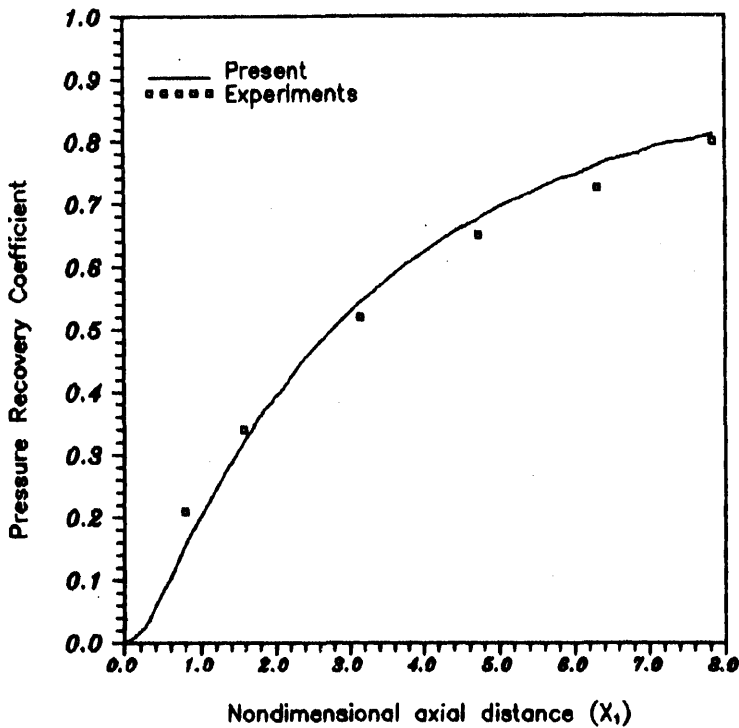


Figure 21. Variation of pressure recovery coefficient for flow through diffuser without inlet velocity distortion.



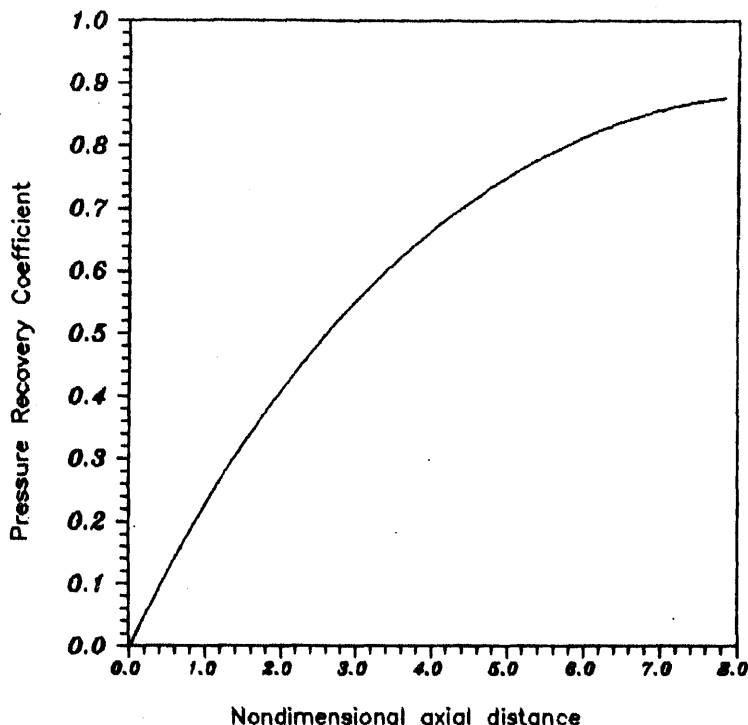
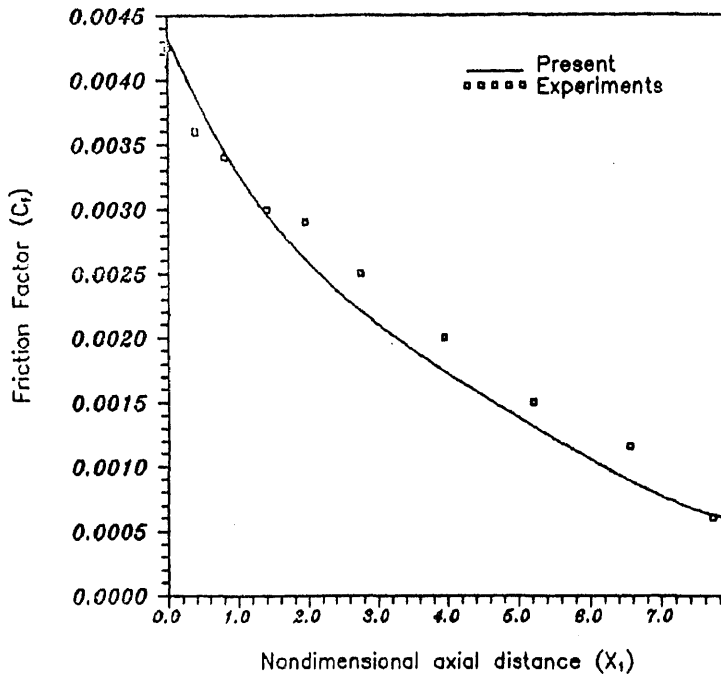


Figure 22. Comparison of pressure recovery coefficient for flow through diffuser with inlet velocity distortion.

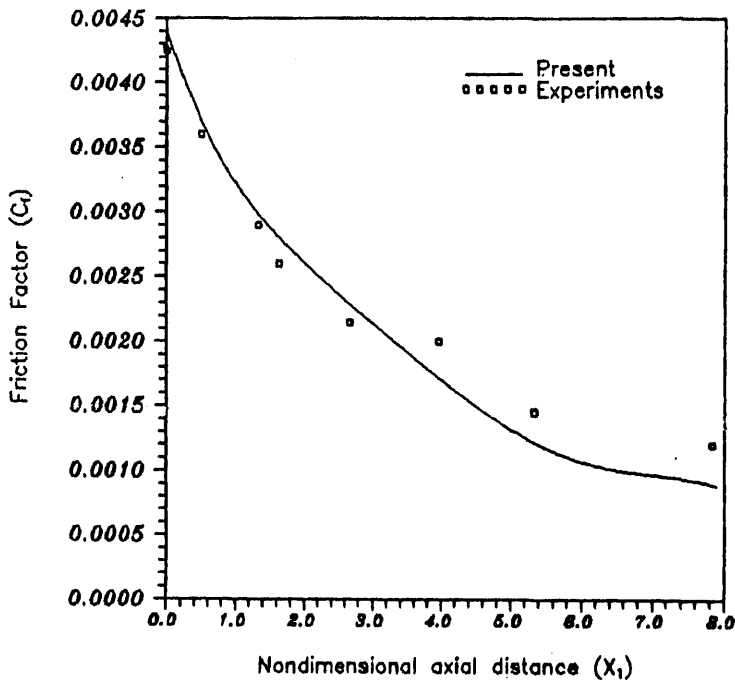
is well developed with parabolic procedures. Patankar and Spalding (Firziger 1987) presented an accurate and economical parabolic method for calculating heat and mass, and momentum transfer in three-dimensional flows. A finite element procedure for the prediction of laminar forced convection in three-dimensional parabolic flows is presented by Nonino *et al* (1988). Local Nusselt number variations are presented and the results are compared with the correlations. Godbole (1975) used a penalty function approach for solving flow at a very low Re number.

In the present study, the complete Navier–Stokes equations are solved. Flow in rectangular ducts is solved for aspect ratios  $\alpha^* = 0.5, 0.75, 1.0$ . Aspect ratio 1.0 is for a square duct. Heat is treated as a scalar in forced convection problems. The temperature field has no influence on the flow field. The Navier–Stokes' and the energy equations are solved as an uncoupled system.

(i) *Geometry and boundary conditions* – Due to symmetry, only a quarter of the duct is solved. All the lengths are non-dimensionalised with the hydraulic diameter of the duct. No-slip boundary conditions are assumed on the walls, symmetric conditions at the axis and also that there should not be any flow across the axis plane. Hence the flow velocity perpendicular to the axis plane is specified as zero on the axis. Traction-free boundary conditions are specified at the exit. Pressure is specified as zero at the exit. Inlet velocity is specified as unity. The hydrodynamic developing length for ducts is approximately 0.01 times the Reynolds number for the square and rectangular ducts. Different lengths are taken for different Re to save computing time. However in the case of the rectangular ducts, the same length is taken for all the "



**Figure 23.** Comparison of friction factor for flow through diffuser without inlet velocity distortion.



**Figure 24.** Comparison of friction factor for flow through diffuser with inlet velocity distortion.

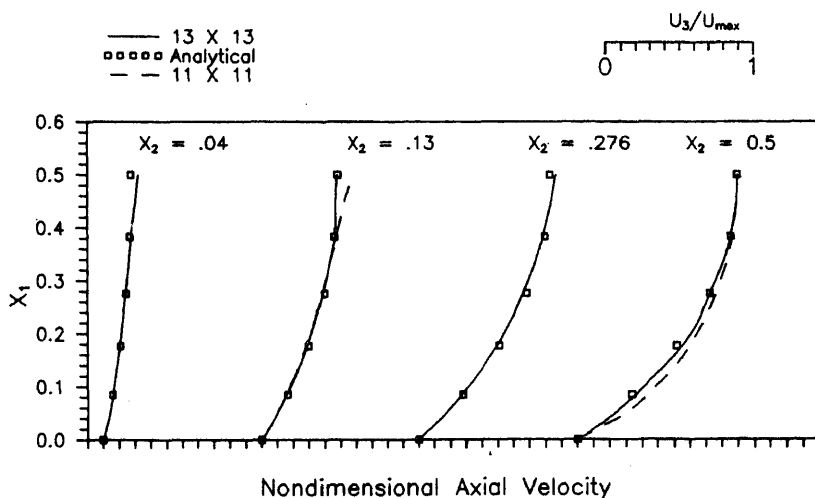
numbers solved i.e. 50, 100 and 200.

The domain is divided into tetrahedral elements. Initially the domain is divided into cuboids. Each cuboid can be divided into either five or six tetrahedral elements without introducing any new nodes. However if the cuboid is divided into five tetrahedra, each cuboid has to be divided in a different manner. In order to avoid this, each cuboid is divided into six tetrahedra. This does not increase the memory requirement since no new nodes are introduced. Also the increase in computational time is negligible.

(ii) *Results and discussion* – Figure 25 shows the comparison with experiments of velocity profiles of square duct. Grid independence tests were conducted with  $11 \times 11$ ,  $13 \times 13$  and  $15 \times 15$  grids. The variation from  $13 \times 13$  in the velocity profile for  $15 \times 15$  is very little. Unlike in 2-D flows, the size of the mesh goes up rapidly even by increasing the mesh size by one node in each section. The bandwidth also goes up rapidly. In order to keep the size of the problem and the computational time required at a reasonable level all the further calculations are carried out with a grid size of  $13 \times 13$ .

Aspect ratio of elements in each section is approximately adjusted to one. The mesh is irregular and at least three nodes are placed within the boundary layer. Also it is observed that it is better to refine the mesh in each section rather than to have more sections in the flow direction for better prediction of fully developed velocity profile, Nusselt number and friction factor. It is also observed that as the mesh is refined the development of the flow is slow. Table 6 shows the comparison of  $U_{\max}/U_m$  and the hydrodynamic length for square ducts for different values of Re. The hydrodynamic developing lengths for square ducts given by Srinivas (1994) are in good agreement with experiments.

Even though the cell Peclet number is larger, stability is taken care of by the time step. All the convective and diffusive terms of Navier–Stokes equations are multiplied with the time step. The maximum time step permissible varies from mesh to mesh. As the mesh becomes finer the time step allowable becomes smaller.



**Figure 25.** Comparison of fully developed axial velocity profiles for flows through square ducts, at different locations.

**Table 6.** Comparison of maximum velocity and hydrodynamic length for flow through a square duct for different values of Re.

Re	$U_{\max}/U_m$	$L_{hy}$		
	Present <sup>†</sup> work	Present work	Dalton*	Han*
50	2.11	4.90	4.5	3.75
100	2.03	10.10	9.0	7.50
200	1.97	19.60	18.0	15.0
500	1.96	43.10	45.0	37.50

<sup>†</sup> Compared with the fully developed value 2.090

\* Shah & London (1978)

The wall derivatives are evaluated using curve fitting techniques. It is observed that 4-point fit is adequate to predict the Nusselt numbers and friction factors. It is not advisable to go beyond 5-point fit, as the higher order polynomial fits are well known for their lack of accuracy and their tendency to manifest kinks and oscillations in the fitted curve. The derivatives are evaluated at all the boundary nodes and the curve is integrated for the evaluation of Nusselt number and friction factor.

Table 7 shows the prediction of fully developed Nusselt numbers for constant temperature boundary condition for various Re and for various  $\alpha^*$ .

Figures 26 and 27 show the variation of non-dimensionalised pressure drop for different Re for  $\alpha^* = 0.5$  and 0.75. It can be seen that no chequer-board oscillations are present in the pressure field prediction.

The skin friction coefficient and the Nusselt number are often used to present the results in a concise manner. Figure 28 shows the variation of friction factor for square ducts for different Re. Figures 29 and 30 show the variation of bulk temperature and Nusselt number for different Pr numbers. The maximum error in the prediction of Nusselt number and friction factor is about 3%. Figure 31 shows the isotherms in fully developed section.

(iii) *Conclusions* – The problems solved reveal many aspects of the developed tool. The results indicate that the present predictions are good. The pressure field shows no chequer-board splitting. A four-point fit is adequate to calculate the wall derivatives.

**Table 7.** Comparison of Nusselt numbers for different Re for square ducts and rectangular ducts.

S. No	Re	Nusselt number		
		$\alpha^* = 1.0$	0.50	0.75
1	50	2.980	3.58	3.11
2	100	2.967	3.62	3.07
3	200	2.910	3.64	3.05

Fully developed value for  $\alpha^* = 1.00$  is 2.996, 0.75 is 3.140, 0.50 is 3.91 (as given in Shah & London 1978)

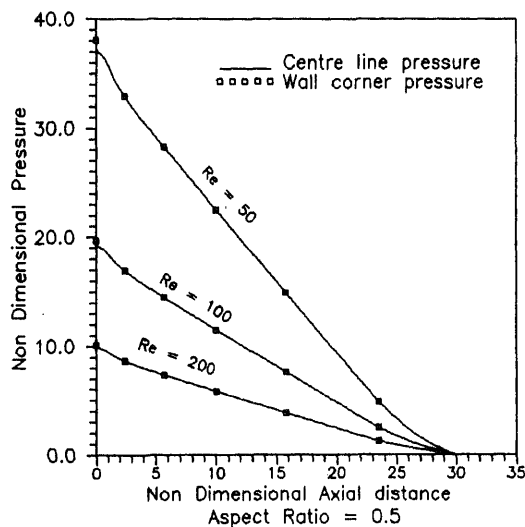


Figure 26. Pressure drop in flow through rectangular duct for different  $Re$  for  $\alpha^* = 0.5$ .

## 5. Closure

The present method offers a cost and storage effective algorithm for prediction of 2-D and 3-D internal flows. It requires a minimum of CPU time and storage space, as no large matrix is being solved. During the early period of the transient behaviour, the flow field and the thermal field evolve rapidly and the time step has to be very small. As time progresses, the rate of convergence becomes slower, i.e., the difference between the values of any field variable for two successive iterations become smaller, at which state the time step can be progressively increased. Thus a judicious selection of time step effects a significant reduction in storage requirement and CPU time. At the same time, accuracies comparable to finite difference methods are achieved. All the present results have been obtained by running the finite element code on an AT 486 machine.

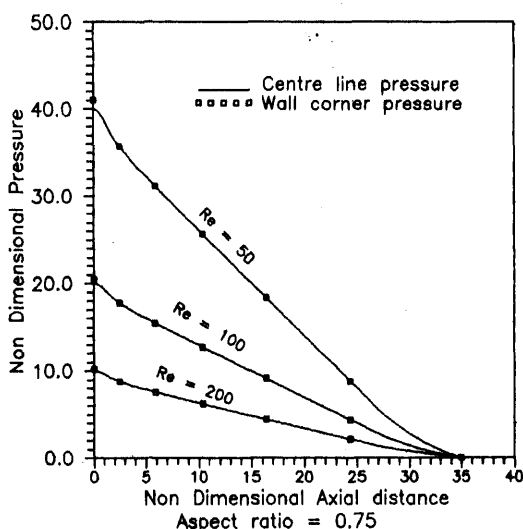


Figure 27. Pressure drop in flows through rectangular ducts for different  $Re$  for  $\alpha^* = 0.75$ .

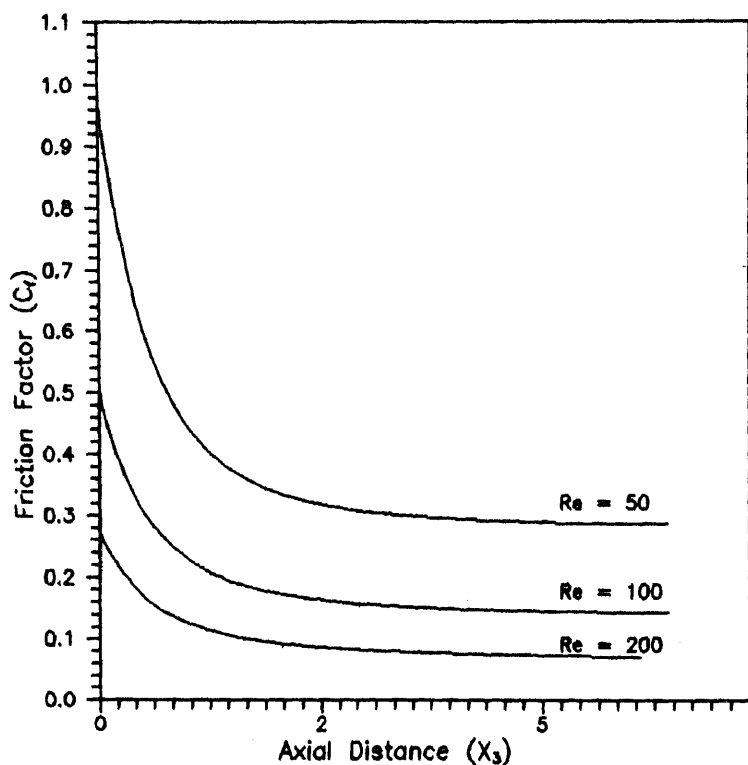


Figure 28. Variation of friction factor for flows through square ducts for different  $Re$ .

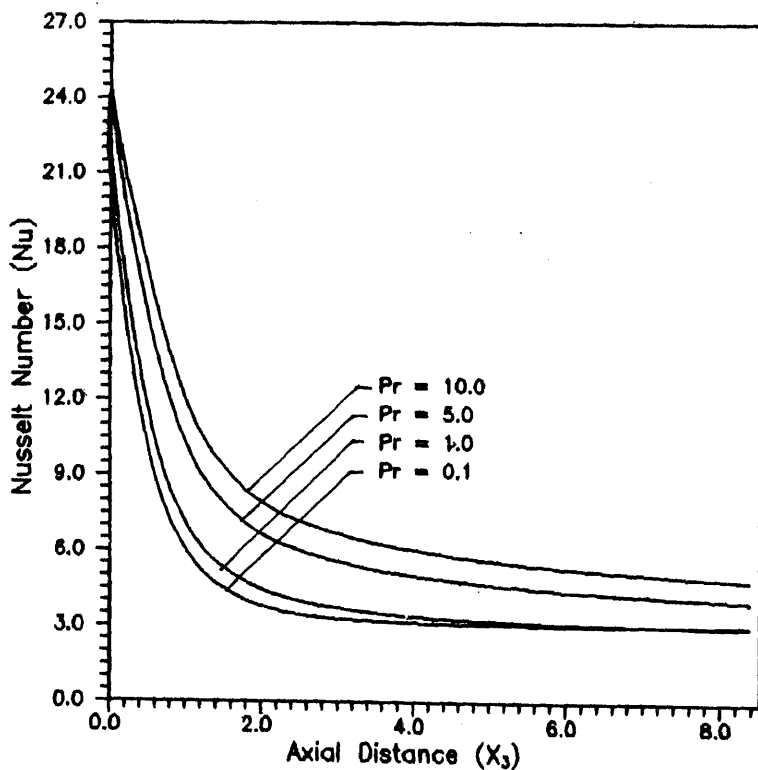


Figure 29. Variation of Nusselt number for flow through square duct for different  $Pr$ .

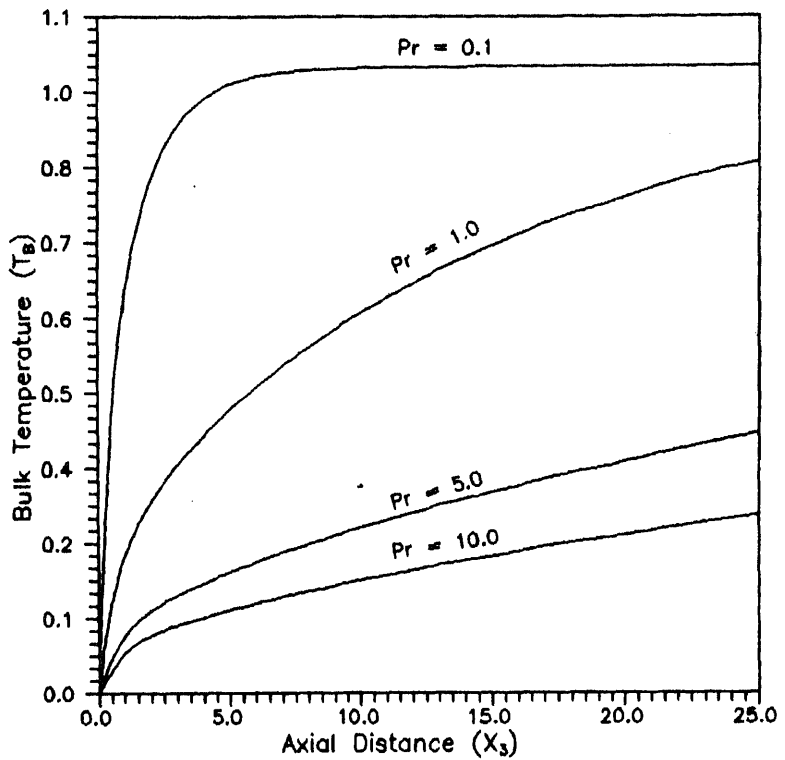


Figure 30. Variation of bulk temperature flow through square duct for different  $Pr$ .

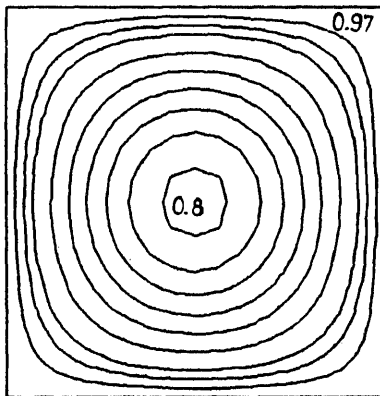


Figure 31. Isothermal plot in an axial cross-section of square duct.

### List of symbols

$C_f$	friction factor;
$C_\mu, C_{\varepsilon 1}, C_{\varepsilon 2}$	turbulence model constants;
$f_\mu, f_1, f_2$	turbulence model functions;
$f$	frequency;

HR	high Reynolds version of turbulence model;
$k$	turbulence kinetic energy
$l$	mixing length;
$L_{ref}$	Reference length for non-dimensionalization;
LR	low Reynolds version of turbulence model;
$n, n + 1$	at $n$ th and $(n + 1)$ th time step;
$Nu$	Nusselt number;
$\bar{Nu}$	average Nusselt number for ribbed channel;
$Nu_s$	Nusselt number of smooth channel;
NML	Nikuradse mixing length;
$P$	mean component of pressure;
$\Delta P$	pressure drop per unit length of ribbed length;
$\Delta P_s$	pressure drop per unit length of smooth channel;
PML	Prandtl mixing length;
$P_i$	pitch of the ribs as shown;
Pe	Peclet number = $Re \cdot Pr$ ;
Pr	Prandtl number;
PR	pitch ratio, $P_i/H$ ;
Re	Reynolds number;
$t$	dimensionless time;
$\Delta t$	time step;
$T_*$	Wall heat flux temperature (used in high Re versions);
$U_1$	flow direction velocity for 2-D problems;
$U_3$	flow direction velocity for 3-D problems;
$U_{ref}$	reference velocity;
$U_*$	friction velocity (used in high Re version);
$V$	pseudo velocity in Eulerian velocity correction method;
VEV	Van Driest effective viscosity model;
$X_1$	flow direction for 2-D problems;
$X_3$	flow direction for 3-D problems;
$X_{R1}$	non-dimensional reattachment length (with rib height) after first rib;
$X_{R2}$	non-dimensional reattachment length (with rib height) after second rib;
$y$	normal wall distance;
$\alpha$	thermal diffusivity;
$\alpha_T$	turbulent thermal diffusivity;
$\varepsilon$	turbulent kinetic energy dissipation;
$\varepsilon_{in}$	inlet value of $\varepsilon$ ;
$\nu$	momentum diffusivity;
$\nu_T$	turbulent momentum diffusivity;
$\eta$	heat transfer augmentation efficiency;
$\rho$	density;
$\sigma_k, \sigma_\varepsilon$	turbulence model constants for diffusion of $k$ and $\varepsilon$ ;
$\theta$	rib wall inclination.

## References

- Acharya S, Dutta S, Myrum A, Baker R S 1993 Periodically developed flow and heat transfer in a ribbed duct. *Int. J. Heat Mass Transfer* 36: 2069–2082



- Autret A, Grandotto M, Dekeyser I 1987 Finite element computation of a turbulent flow over a two dimensional backward facing step. *Int. J. Numer. Methods Fluids* 7: 89–102
- Benim A C, Zinser W 1985 Investigation into the finite element analysis of confined turbulent flows using a  $k$ - $\epsilon$  model. *Comput. Methods Appl. Mech.* 51: 507–523
- Bradshaw P *et al* 1981 *Engineering calculation methods for turbulent flows* (London: Academic Press)
- Caretto L S *et al* 1972 Two numerical methods for 3-D boundary layers. *Comput. Methods Appl. Mech. Eng.* 1: 39–57
- Chitambar V C 1978 *Some studies on the structure of incompressible turbulent flow in a two-dimensional diffuser with inlet velocity distortions*, PhD thesis, Indian Institute of Technology, Madras
- Curr R M, Sharma D, Tatchell 1972 Numerical predictions of some 3-D boundary layers in ducts *Comput. Methods Appl. Mech. Eng.* 1: 143–158
- Deissler R G 1988 Turbulent solutions of the Navier–Stokes equations. *Encyclopedia of fluid mechanics* (Houston, TX: Gulf) vol. 6, chap. 29
- Donea J, Ginliani S, Laval H, Quartapele 1982 Finite element solution of unsteady Navier–Stokes equations by fractional step method. *Comput. Methods Appl. Mech. Eng.* 30: 53–73
- Ferziger J H 1987 Simulation of incompressible turbulent flows. *J. Comput. Phys.* 69: 1–48
- Gresho P M, Sani R L 1987 On pressure boundary conditions for the incompressible Navier–Stokes equations. *Int. J. Numer. Methods Fluids* 7: 1111–1145
- Gresho P M *et al* 1981 Solution of the time dependent 3-D incompressible Navier–Stokes equation via FEM. In *Proc. Int. Conf. on Numerical Methods in Laminar and Turbulent flows* (eds) C Taylor *et al*, UK
- Godbole P N 1975 Creeping flow in rectangular ducts by finite element method. *Int. J. Numer. Methods Eng.* 727–731
- Hah C 1983 Calculation of various diffuser flows with inlet swirl and inlet distortions effects. *AIAA J.* 21: 1127–1133
- Haines A B 1982 Turbulence modelling – Report of a working party. *Aeronaut. J.* 86: 269–277
- Han J C, Glicksman L R, Roshenow W M 1978 An investigation of heat transfer and friction factor for rib roughened surface. *Int. J. Heat Mass Transfer* 27: 1143–1155
- Hoffmann J A 1982 Effects of free stream turbulence on diffuser performance. *J. Fluids Eng., ASME* 103: 104–110
- Hussaini M Y, Zang T A 1987 Spectral methods in fluid dynamics. *Annu. Rev. Fluid Mech.* 19: 339–367
- Hutton A G, Smith R M 1987 The computation of turbulent flows of industrial complexity by finite element method – Progress and prospects. In *Finite elements in fluids* (eds) R H Gallagher *et al* (New York: John Wiley) 7: 289–310
- Johnson R W, Launder B E 1982 Discussions on the calculation of turbulent heat transport down stream of an abrupt pipe expansion. *Numer. Heat Transfer* 5: 493–495
- Lakshminarayana B, Kirtley K R, Warfield M 1989 Computational techniques and validation of 3-dimensional viscous turbulent codes for internal flows. *Sādhanā* 14: 59–91
- Lam C K G, Bremhorst K 1981 A modified form of  $k$  –  $\epsilon$  model for predicting wall turbulence. *J. Fluids Eng.* 103: 456–460
- Launder B E 1982 A generalized algebraic stress transport modelling. *AIAA J.* 20: 436–437
- Launder B E 1984 Numerical computation of convective heat transfer in complex turbulent flows: Time to abandon wall functions. *Int. J. Heat Mass Transfer* 27: 1485–1491
- Launder B E 1988 On the computation of convective heat transfer in complex turbulent flow. *J. Heat Transfer* 110, 1112–1128.
- Liou T M, Chang Y, Huang D W 1990 Experimental and computational study of turbulent flows in a channel with two pairs of turbulence promoters in tandem. *J. Fluids Eng.* 112: 302–310
- Lumley J L 1983 Turbulence modelling. *ASME J. Appl. Mech.* 50: 1097–1103
- Markatos M C 1988 Computer simulation techniques for turbulent flows. In *Encyclopedia of fluid mechanics* (Houston, TX: Gulf) vol. 6, chap. 28
- Martinuzzi R, Pollard A 1989 Comparative study of turbulence models in predicting turbulent pipe flow. Part I: Algebraic stress and  $k$  –  $\epsilon$  models. *AIAA J.* 27: 29–36
- Mellor G L, Herring J H 1973 A survey of mean turbulent filled closure models. *AIAA J.* 11: 590–599

- Michelic M, Wingerath K 1988 Numerical solution of free and forced convection bulk flows. In *Encyclopedia of fluid mechanics* (Houston, TX: Gulf) vol. 6, chap. 35
- Morgan K et al 1987 Investigation of mixing length and two-equation turbulence models utilizing the finite element method. *Appl. Math. Modelling* 1: 395–400
- Murphy J D 1988 Turbulence modelling. In *Encyclopedia of fluid mechanics* (USA: Gulf) vol. 6, chap. 28
- Nallasamy M 1987 Turbulence models and their applications to the prediction of internal flows: A review. *Comput. Fluids* 15: 151–194
- Nonino C, Del Guidice, Comini G 1988 Laminar forced convection in 3-D duct flows. *Numer. Heat Transfer* 13: 451–466
- Patankar S V 1988 Recent developments in computational heat transfer. *ASME J. Heat Transfer* 110: 1037–1045
- Pletcher R H 1988 Progress in turbulent forced convection. *ASME J. Heat Transfer* 110: 1129–1144
- Rabas T J 1989 Selection of energy efficient enhancement geometry for single phase turbulent flows inside tubes. *ASME Proc. of 1989 Natl. Heat Conf., HTD* vol. 108, pp. 193–204
- Ravikumaur S G 1988 *Finite element analysis of convective heat transfer and heat exchangers*. Ph D thesis, Indian Institute of Technology, Madras
- Ravisankar M S 1992 *Finite element analysis of turbulent flows with heat transfer*. M S thesis, Indian Institute of Technology, Madras
- Reddy J N 1982 Penalty finite element analysis of 3-D Navier–Stokes equations. *Comput. Methods Appl. Mech. Eng.* 35: 87–106
- Reynolds W C 1978 On calculation of turbulent flows. In *Turbulence* (ed.) P Bradshaw (New York: Springer-Verlag) chap. 5
- Rodi W 1982 Examples of turbulent flow in smooth pipe. *Appl. Sci. Res.* 28: 872–879
- Rodi W 1984 Turbulence models and their applications to hydraulics – State of art review. *Intl. Assos. Hydraulic Res., Delft*
- Shah R K, London A L 1978 Laminar flow forced convection in ducts. *Advances in heat transfer: Supplement 1* (New York: Academic Press)
- Shih T M 1985 A literature survey on numerical heat transfer. *Numer. Heat Transfer* 8: 1–24
- Shih T M 1987 A literature survey on numerical heat transfer. *Numer. Heat Transfer* 11: 1–29
- Shih T M 1989 A literature survey on numerical heat transfer. *Numer. Heat Transfer* 15: 1–39
- Smith R M 1984 A practical method of two-equation modelling using finite elements. *Int. J. Numer. Methods Fluids* 4: 321–336
- Srinivas M 1994 *Finite element analysis of internal flows with heat transfer*. Ph D thesis, Indian Institute of Technology Madras
- Taylor C et al 1977 A numerical analysis of turbulent flow in pipes. *Comput. Fluids* 5: 191–203
- Taylor C, Harper J J, Hughes T G, Morgan K 1981 An analysis of developing turbulent flow in a circular pipe by the finite element method. In *Proc. Numer. Methods Laminar and Turbulent flows* (eds) Baker et al (Swansea: Pineridge)
- Unes T 1988 Two-equation ( $k - \epsilon$ ) turbulence computations by the use of a finite element model. *Int. J. Numer. Methods Fluids* 8: 965–975

## Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials

S GOPALAKRISHNA and Y JALURIA

Department of Mechanical and Aerospace Engineering, Rutgers University,  
New Brunswick, NJ 08903, USA

**Abstract.** A numerical study of the transport phenomena arising in a single-screw extruder channel is carried out. A non-Newtonian fluid is considered, using a power law model for the variable viscosity. Chemical reaction kinetics are also included. Finite difference computations are carried out to solve the governing set of partial differential equations for the velocity, temperature and species concentration fields, over a wide range of governing parameters for the case of a tapered screw channel.

The numerical treatment for this combined heat and mass transfer problem is outlined. A marching procedure in the down-channel direction is adopted and the validity of the scheme for practical problems discussed. For large viscous dissipation, the material heats up considerably due to the prevailing shear field, affecting the viscosity significantly, and results in large changes in the pressure development at the end of the channel. The rate of reaction controls the mass diffusion rate which in turn affects viscosity and the flow significantly. The dimensionless throughput,  $q_v$ , is one of the most important parameters in the numerical solution. The dimensionless pressure variation is very sensitive to  $q_v$ , and orders of magnitude changes are possible for small variations in  $q_v$ . Schemes for dealing with other important effects such as back flow, heat transfer by conduction in the barrel, and the effect of the die are also outlined.

**Keywords.** Transport phenomena; single-screw extruder; chemical reaction kinetics; finite difference computations; non-Newtonian materials.

### 1. Introduction

Screw extrusion is a thermomechanical processing operation in which the raw material is fed into a hopper and forced through the passage between a rotating screw and a stationary barrel. The processed material comes out through a die of a specific shape. Single- and twin-screw extruders are used widely in the food and plastics processing industry for the production of shaped and cooked products. The high shear and temperature environment inside the screw channel results in mixing of the material

and leads to chemical reactions that constitute the cooking process. The underlying heat and mass transport processes have been analysed, but a comprehensive treatment that is useful for extruder design and optimization is not available. This chapter attempts to fill the gap in the literature by numerically simulating the complex heat and mass transfer interactions for the simple geometry of the single-screw extruder.

Several researchers have studied the flow of polymers in the various sections of an extruder, using different numerical or analytical techniques (Fenner 1977, 1979; Tadmor & Gogos 1979). Fenner (1977) considered the case of the temperature profile developing along the length of the screw channel. Elbirli & Lindt (1984) have reported the results from a model in which the temperature was allowed to develop along the screw channel. In these models, the screw and the barrel were assumed to be at the same uniform temperature. Karwe & Jaluria (1990) have presented numerical results for flow and heat transfer for polymeric materials in single-screw extruders with an adiabatic boundary condition at the screw. There is no available literature on the simulation of mass transport in screw extrusion, even though it is of primary practical importance in terms of product quality and attributes of the extrudate.

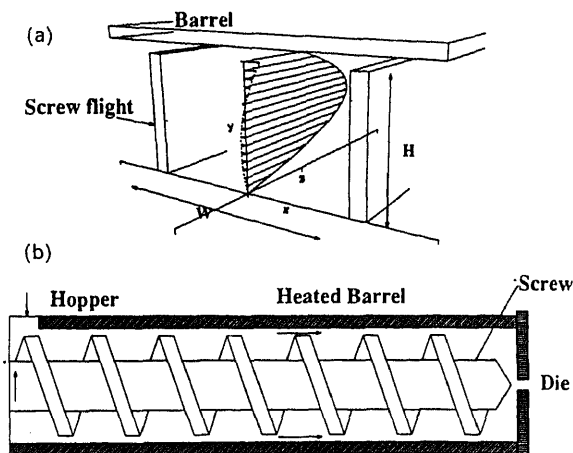
A finite difference numerical study has been carried out here for the simulation of flow and transport processes in extrusion of food and plastic materials, to obtain the velocity, temperature, and mass concentration variation along the length of the screw channel. Starch granules are gelatinized by means of water absorption inside the extruder. This conversion process is modelled in our analysis. The variation in channel depth due to taper is accounted for in the model. The effects of viscous dissipation and chemical reaction are included, and the results obtained are discussed in terms of the basic transport processes. The main governing dimensionless parameters are the taper angle  $\psi$ , the throughput, or volume flow rate  $q_v$ , the Peclet number  $Pe$ , which represents the relative importance of convection along the screw channel as compared to conduction, the power law index  $n$ , the dimensionless viscosity coefficients  $\beta, \beta_1$ , and  $b_m$ , which characterize the variation of viscosity with temperature and concentration, the dimensionless gelatinization rate  $S$ , and the Griffith number  $G$ , which represents the level of viscous dissipation compared to heat conduction. The analysis applies to the metering section or the melt conveying section, where the material is heated and subjected to high shear rates. The effect of the governing parameters on the flow, heat transfer, and concentration distribution within the screw channel is studied in detail. It is found that the pressure developed at the die is strongly affected by the throughput, the reaction rate  $S$  and the power law index.

## **2. Problem formulation**

The simplified geometry of a single-screw extruder and the cross section of a screw channel are shown in figure 1. For ease of visualization and analysis, the coordinate system is fixed to the screw root and, thus, the barrel moves in a direction opposite to the screw rotation. Such a formulation is commonly employed in the literature (Fenner 1979; Tadmor & Gogos 1979).

The following assumptions are made in deriving the governing equations from the basic conservation principles.

- (1) Curvature of the screw channel is small, enabling the channel helix to be "opened out" to obtain a simpler geometry for ease of computation.



**Figure 1.** Simplified geometry of a single-screw extruder with a rectangular screw profile. (a) Perspective and (b) cross-sectional views.

- (2) The screw profile is rectangular, the width ( $W$ ) being much larger than the depth ( $H$ ). Other shapes such as a trapezoid can also be simulated in this scheme by using an "equivalent" height.
- (3) The lubrication approximation is valid. For the highly viscous flow regime considered here for typical food materials, the Reynolds number is of the order of 0.0002, showing that the inertial terms are indeed negligible in comparison to the viscous terms.
- (4) Back flow in the channel cross-section is limited to a very small fraction of the flow rate. This assumption relates the presence of a die at the end of the channel to the flow characteristics within the extruder; this coupling is discussed in detail later.
- (5) The clearance between the screw flight and the barrel is assumed to be enough to permit one to neglect the leakage across the flights from one screw channel to the neighbouring one.
- (6) The diffusion of moisture into the food material and its absorption due to gelatinization is modelled as a zeroth-order reaction.

For steady, developing, two-dimensional flow of a homogeneous fluid in a single-screw extruder with a shallow channel (i.e., for  $H \ll W$  in figure 1), after neglecting the inertia terms (creeping flow approximation: Schlichting 1979) in the  $x$  and  $z$  directions, the equations for the conservation of momentum become:

$$\frac{\partial p}{\partial x} = \frac{\partial(\tau_{yx})}{\partial y}, \quad \frac{\partial p}{\partial y} = 0, \quad \frac{\partial p}{\partial z} = \frac{\partial(\tau_{yz})}{\partial y}, \quad (1)$$

$$\tau_{yx} = \mu(\partial u / \partial y), \quad \tau_{yz} = \mu(\partial w / \partial y). \quad (2)$$

The energy equation, after employing the lubrication approximation, becomes (Kays & Crawford 1980):

$$\rho C w \frac{\partial T}{\partial z} = \frac{\partial}{\partial y} \left( k \frac{\partial T}{\partial y} \right) + \tau_{yx} \frac{\partial u}{\partial y} + \tau_{yz} \frac{\partial w}{\partial y}. \quad (3)$$

Mathematically speaking, the energy equation is parabolic in the  $z$ -direction. In actual practice, there is a strong restriction to the flow in the form of a die at the end of the extruder. The characteristics of the flow in the extruder are strongly coupled

to that inside the die. For small throughputs, there is back flow in the extruder channel, and this makes the problem elliptic. In such circumstances, the fluid flow and the heat transfer at the die also need to be simulated and coupled to the extruder simulation. The present study is, however, restricted to the simulation of the transport phenomena in the extruder channel. The parabolic nature of the governing equation in the  $z$ -direction allows us to use a marching scheme, as described in the next section.

The constitutive equation for viscosity of starch materials such as Amioca and Hylon 7 is written as:

$$\mu = \mu_0(\dot{\gamma}/\dot{\gamma}_0)^{(n-1)} \exp(b/T) \exp(-B_m C_m), \quad (4)$$

where

$$\dot{\gamma} = [(\partial u/\partial y)^2 + (\partial w/\partial y)^2]^{1/2}.$$

The dependence of viscosity on temperature and moisture content described above has been confirmed by rheological experiments carried out by Kokini *et al* (1987).

The mass diffusion equation also needs to be considered in addition to the continuity, momentum, and energy equations for food systems. The particular food systems we are interested in examining are starches of various types. For example, some of the important effects of moisture transport are manifested in the form of gelatinization of the granular food material, which is fed into the hopper of the screw extruder. This reaction occurs between the starch granules and the water molecules. In this study, for the modelling of moisture diffusion, gelatinization is defined as the process by which water is bonded to the starch granules in the food material and thus becomes unavailable for diffusion. The diffusivity of bound moisture is reduced by several orders of magnitude, and the food material is said to undergo a form of transition. This transition affects the heat absorbed as the material undergoes chemical changes during its passage along the screw channel. The moisture diffuses, gets absorbed, and is convected along the flow direction. In this study, the transport equation for moisture is written to include the effect of gelatinization, and the chosen constitutive equation for viscosity reflects the experimentally observed dependence of viscosity on moisture concentration. It has been observed (Harper 1980) that even small changes in the moisture content can result in large changes in viscosity. The mass transfer equation is written as:

$$w \frac{\partial c_m}{\partial z} = \frac{\partial^2 c_m}{\partial y^2} + S'(c_m)^m, \quad S' = 0 \quad \text{for} \quad T < T_{\text{gel}}. \quad (5)$$

The last term in the right-hand side of the equation represents a source/sink for the diffusing species due to reaction. This term becomes operative only when the temperature exceeds the temperature needed for the onset of the gelatinization reaction,  $T_{\text{gel}}$ . A zero-order reaction is considered in this study, with  $m = 0$ . Wang *et al* (1989) have examined the rates of conversion of starch materials based on shear and thermal processing history. The proportionality constant is a function of the temperature and the shear rate, but is taken as constant in this study as a first approximation, and a parametric study is carried out.

The boundary conditions are prescribed as shown in figure 2. The screw has been taken as isothermal, at  $T_b$ , in most studies reported in the literature. However, a more practical circumstance is represented by the adiabatic condition at the screw surface (Karwe & Jaluria 1990). Otherwise, a conjugate problem needs to be examined, where the conduction in the screw is coupled to the conduction and convection inside the flow channel.

Two constraints arise on the basis of flow rate conservation considerations. If the

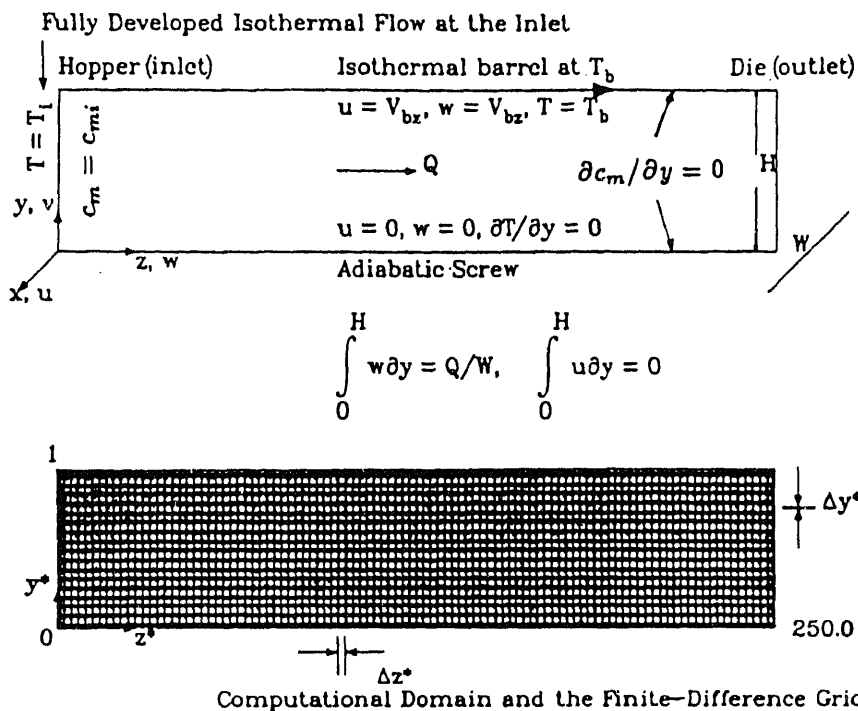


Figure 2. Boundary conditions for the numerical model, employing a coordinate system fixed to the screw.

leakage flow across the screw flights is taken as negligible, these constraints are given, for a total volumetric flow rate  $Q$ , by

$$\int_0^H u \, dy = 0, \quad \text{and} \quad \int_0^H w \, dy = Q/W. \quad (6)$$

For tapered channels, the linear variation of channel height with distance along the screw helix can be written as:

$$H(z) = H_1 - (z/z_{\max})(H_1 - H_2).$$

The governing equations are nondimensionalized in terms of the following dimensionless variables:

$$\begin{aligned} x^* &= x/H_2, \quad y^* = y/H, \quad z^* = z/H_2, \\ u^* &= u/V_{bz}, \quad w^* = w/V_{bz}, \\ \theta &= (T - T_i)/(T_b - T_i), \quad c^* = c_m/c_{mi}, \\ \beta &= T_b/T_i, \quad \beta_1 = b/T_i, \quad b_m = B_m c_{mi}, \\ \dot{\gamma}^* &= \dot{\gamma}H/V_{bz}, \quad \bar{\mu} = \mu_0 [(V_{bz}/H_2)/\dot{\gamma}_0]^{n-1}, \\ p^* &= p/\bar{p}, \quad \bar{p} = \bar{\mu} (V_{bz}/H_2), \quad V_{bz} = \pi DN (\cos \phi/60), \\ \text{Pe} &= V_{bz} H_2/\alpha, \quad G = \bar{\mu} V_{bz}^2/k(T_b - T_i), \\ \text{Le} &= D/\alpha, \quad S = S' H_2^2 c_{mi}^{m-1}/\alpha. \end{aligned} \quad (7)$$

The resulting dimensionless equations are:

$$\frac{\partial p^*}{\partial x^*} = \left(\frac{H_2}{H}\right)^{n+1} \frac{\partial}{\partial y^*} \left\{ \frac{\partial u^*}{\partial y^*} [\dot{\gamma}^*]^{(n-1)} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*) \right\}, \quad (8)$$

$$\frac{\partial p^*}{\partial z^*} = \left(\frac{H_2}{H}\right)^{n+1} \frac{\partial}{\partial y^*} \left\{ \frac{\partial w^*}{\partial y^*} [\dot{\gamma}^*]^{(n-1)} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*) \right\}, \quad (9)$$

$$\begin{aligned} \text{Pe } w^* \frac{\partial \theta}{\partial z^*} &= \left(\frac{H_2}{H}\right)^2 \frac{\partial^2 \theta}{\partial y^{*2}} \\ &+ \left(\frac{H_2}{H}\right)^{n+1} G [\dot{\gamma}^*]^{(n+1)} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*), \end{aligned} \quad (10)$$

$$\text{Pe } w^* \frac{\partial c^*}{\partial z^*} = \text{Le} \left(\frac{H_2}{H}\right)^2 \frac{\partial^2 c^*}{\partial y^{*2}} + S c^{*n}, \quad S = 0 \quad \text{for } \theta < \theta_{\text{gel}}. \quad (11)$$

Similarly, the boundary conditions are also obtained in the dimensionless form. The constraints on the flow are obtained, in dimensionless form, as follows:

$$\begin{aligned} \int_0^1 u^* dy^* &= 0, \\ \int_0^1 w^* dy^* &= (Q/W)/H V_{bz} = q_v H_2/H, \end{aligned} \quad (12)$$

where

$$q_v = (Q/W)/H_2 V_{bz}.$$

Thus, for a given screw configuration, the parameters that govern the problem are  $\psi, n, \beta, \beta_1, b_m, \theta_{\text{gel}}, \text{Pe}, G, \text{Le}$ , and  $q_v$ .

### 3. Solution procedure

The dimensionless equations are solved by means of finite difference techniques. The computations were carried out over  $y \times z$  grid sizes of  $41 \times 101$ ,  $61 \times 101$ , and  $81 \times 101$ . The results were essentially unchanged when the grid was refined to  $81 \times 121$  from  $61 \times 101$  and therefore, a  $61 \times 101$  grid was selected. Since the energy and mass transport equations, (10) and (11), are parabolic in  $z$ , boundary conditions are necessary only at  $z=0$  to allow marching in the  $z$ -direction and, thus, obtain the solution in the entire domain. The throughput parameter  $q_v$  determines the limits of applicability of the marching scheme. As a rough estimate, a value of  $q_v$  less than 0.2 resulted in significant back flow, for typical values of the other parameters, limiting the use of marching for the numerical solution.

A method similar to the one developed by Fenner (1979) is employed for solving the momentum equations, (8) and (9), at a given  $z$  location iteratively.

In dimensionless form, the momentum equations in  $z$  and  $x$  directions are:

$$\pi_z = \frac{\partial}{\partial z^*} \left( \dot{\gamma}^{*n-1} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*) \frac{\partial w^*}{\partial y^*} \left(\frac{H_2}{H}\right)^{n+1} \right), \quad (13)$$



$$\pi_x = \frac{\partial}{\partial y^*} \left( \dot{\gamma}^{*n-1} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*) \frac{\partial u^*}{\partial y^*} \left( \frac{H_2}{H} \right)^{n+1} \right), \quad (14)$$

where  $\pi_z = \partial p^*/\partial z^*$  and  $\pi_x = \partial p^*/\partial x^*$ .

Integrating (13) and (14) over  $y^*$ , we get:

$$\pi_z(y^* - y_0^*) = \dot{\gamma}^{*n-1} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*) \frac{\partial w^*}{\partial y^*} \left( \frac{H_2}{H} \right)^{n+1}, \quad (15)$$

$$\pi_x(y^* - y_1^*) = \dot{\gamma}^{*n-1} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(-b_m c^*) \frac{\partial u^*}{\partial y^*} \left( \frac{H_2}{H} \right)^{n+1}, \quad (16)$$

where  $y_0^*$  and  $y_1^*$  represent  $y$ -locations in the channel (from the screw root to the barrel) where the shear stresses are zero.

Rearranging, we have

$$\partial w^*/\partial y^* = \pi_z(y^* - y_0^*) F(y^*), \quad (17)$$

$$\partial u^*/\partial y^* = \pi_x(y^* - y_1^*) F(y^*), \quad (18)$$

where

$$F(y^*) = \dot{\gamma}^{*1-n} \exp \left[ \frac{\beta_1}{\theta(\beta-1)+1} \right] \exp(b_m c^*) \left( \frac{H}{H_2} \right)^{n+1}. \quad (19)$$

Squaring and adding (17) and (18), we get:

$$\begin{aligned} (\dot{\gamma}^*)^2 &= (\partial w^*/\partial y^*)^2 + (\partial u^*/\partial y^*)^2, \\ &= [\pi_z^2(y^* - y_0^*)^2 + \pi_x^2(y^* - y_1^*)^2] F^2(y^*). \end{aligned} \quad (20)$$

Solving for  $\dot{\gamma}^*$ , we get:

$$\begin{aligned} \dot{\gamma}^* &= [\pi_z^2(y^* - y_0^*)^2 + \pi_x^2(y^* - y_1^*)^2]^{1/2n} \times \\ &\quad \exp \left[ \frac{-\beta_1/n}{\theta(\beta-1)+1} \right] \exp \left( \frac{b_m c^*}{n} \right) \left( \frac{H}{H_2} \right)^{(n+1)/n} \end{aligned} \quad (21)$$

Substituting back into (19), we write

$$\begin{aligned} F(y^*) &= [\pi_z^2(y^* - y_0^*)^2 + \pi_x^2(y^* - y_1^*)^2]^{(1-n)/2n} \\ &\quad \exp \left[ \frac{-\beta_1/n}{\theta(\beta-1)+1} \right] \exp \left( \frac{b_m c^*}{n} \right) \left( \frac{H}{H_2} \right)^{(n+1)/n}. \end{aligned} \quad (22)$$

The velocities are obtained by integrating (17) and (18):

$$w^* = \int_0^{y^*} \pi_z(\alpha - y_0^*) F(\alpha) d\alpha, \quad (23)$$

$$u^* = \int_0^{y^*} \pi_x(\alpha - y_1^*) F(\alpha) d\alpha. \quad (24)$$

The conditions to be satisfied by the velocities  $u^*$  and  $w^*$  are:

$$\begin{aligned} w^* \{ \text{at } y^* = 1 \} &= 1, \\ u^* \{ \text{at } y^* = 1 \} &= \tan \phi, \\ q_v H_2 / H &= \int_0^1 w^* dy^*, \\ 0 &= \int_0^1 u^* dy^*. \end{aligned} \quad (25)$$

After some algebraic manipulations, the following equations for the four unknowns in the left hand side are obtained:

$$\begin{aligned} \pi_z &= (J_0 - J_1 - J_0 q_v H_2 / H) / (J_0 J_2 - J_1^2), \\ \pi_z y_0^* &= (J_1 - J_2 - J_1 q_v H_2 / H) / (J_0 J_2 - J_1^2), \\ \pi_x &= (J_0 - J_1) \tan \phi / (J_0 J_2 - J_1^2), \\ \pi_x y_1^* &= (J_1 - J_2) \tan \phi / (J_0 J_2 - J_1^2), \end{aligned} \quad (26)$$

where

$$J_m = \int_0^1 \alpha^m F(\alpha) d\alpha. \quad (27)$$

The solution algorithm is enumerated below.

- (1) Guess  $\pi_z, \pi_z y_0^*, \pi_x, \pi_x y_1^*$ .
- (2) Calculate  $F(y^*)$  using (22).
- (3) Calculate the  $J$  integrals using (27).
- (4) Solve (26) using the Gauss-Jordan algorithm.
- (5) Obtain converged flow solution at each time step.
- (6) Using the converged flow solution, march down the channel to obtain temperature and moisture concentration fields at the next downstream location using (10) and (11).
- (7) Go back to step 1 with the new set of temperature and concentration values.

The iterations in the flow calculation are complete when the pressure gradients satisfy the following convergence criterion:

$$\max[\Delta(\partial p^* / \partial z^*), \Delta(\partial p^* / \partial x^*)] \leq 10^{-4}, \quad (28)$$

where  $\Delta$  stands for the absolute value of the fractional change between two consecutive iterations. This particular convergence criterion is not useful when the values of the pressure gradients become very small (i.e., close to zero). Under such circumstances, only the absolute change in the values of the pressure gradient is considered for convergence. Values of the criterion other than  $10^{-4}$  were also tried, and it was found that satisfactory convergence was obtained with the foregoing value within a reasonable number of iterations (typically 5–10).

Using the boundary conditions, in terms of  $u^*$ ,  $w^*$ ,  $\theta$ , and  $c^*$  at any upstream  $z$  location, the energy equation (10), is solved to obtain the temperature distribution at the next downstream  $z$  location. Equation (10) is solved using the fully implicit scheme (Jaluria & Torrance 1986). In this scheme, a tridiagonal system of equations

is obtained by using the new, uncalculated, value of the dependent variable from the differencing operation in the  $y$ -direction at any nodal point in the numerical scheme. The tridiagonal system is solved using the well-known Thomas (TDMA) algorithm, which is very efficient (Jaluria 1988). The mass transport equation, (11), is solved next using, once again, the fully implicit scheme to obtain the values of concentration at the next location in the marching direction. With the temperature and concentration distributions obtained at the next downstream location, the momentum equations, (8) and (9), are solved iteratively as discussed above to obtain the velocity distribution there. This procedure is repeated until the end of the extruder channel is reached. The integration in (27) was carried out numerically using Simpson's one-third rule (Jaluria 1988).

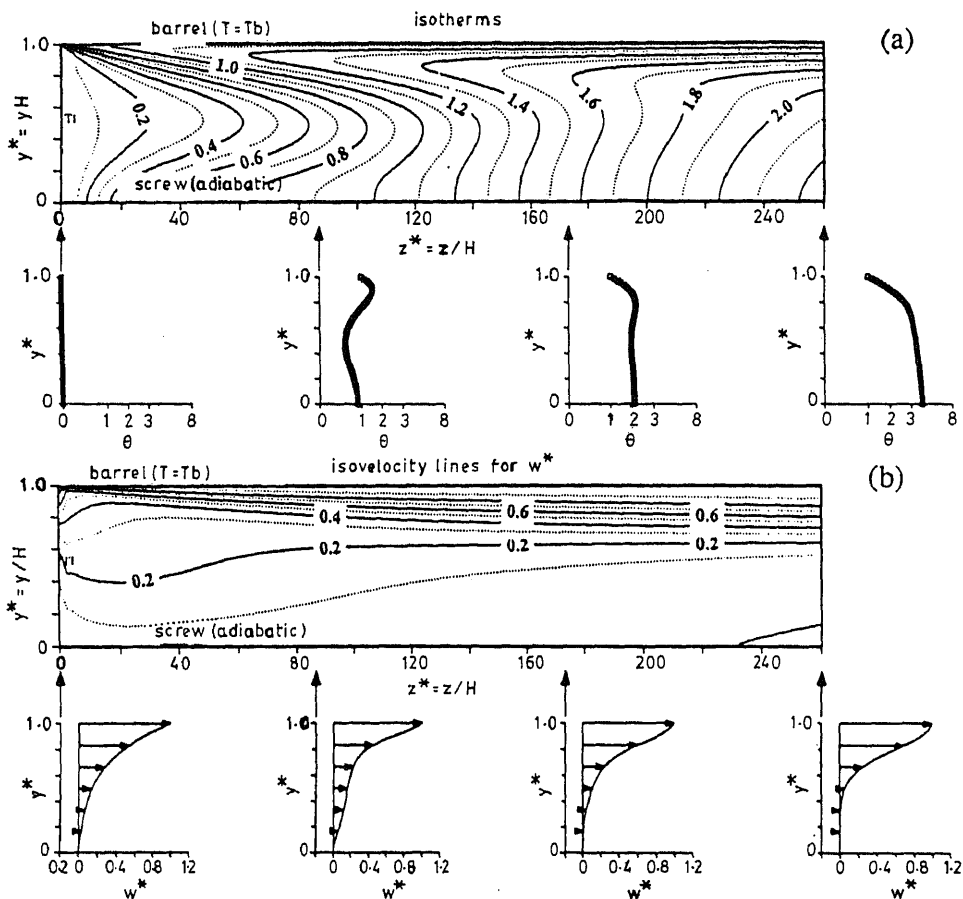
During the course of carrying out the parametric study, a number of interesting features of the numerical scheme were noted. It was found that viscosity characterization in terms of  $\beta$ ,  $\beta_1$ , and  $b_m$  is very closely linked to the magnitude of the gelatinization rate  $S$  that is specified for the particular reaction. For given values of  $\beta$ ,  $\beta_1$ , and  $b_m$ , the pressure development as a function of downstream channel distance is affected considerably by minor changes in  $S$ . In addition, the range of  $q_v$  that produces reasonable pressure development with channel distance is quite limited by back flow considerations on the one hand and unacceptable pressure drops on the other.

#### 4. Results and discussion

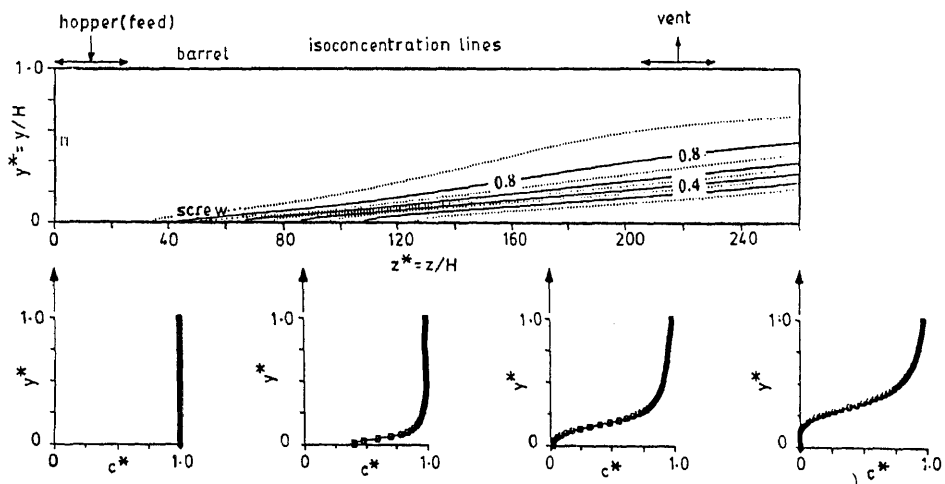
Results are presented in terms of distributions of velocity, temperature, and moisture concentration along the screw channel as well as across the channel depth. The variation of pressure along the extruder channel as a function of the governing parameters is a useful measure of its pumping ability, and is examined here. Contour plots of constant velocity, temperature, and moisture are also obtained. The coordinate system is fixed to the screw as described earlier. For ease of discussion and presentation of results, some of the parameters are kept constant while the others are changed parametrically to study their effect on the transport processes. For the results presented here, the following are kept constant:  $Pe = 7050$ ,  $Le = 0.001$ ,  $\theta_{gel} = 0.5$ ,  $\phi = 16.54^\circ$ ,  $\psi = 0.0^\circ$ , and  $\beta = 1.134$ . In addition, the calculations are carried out for  $\beta_1 = 10$  and  $b_m = 0$ . Other values of these parameters were also considered, but those results are not reported here for brevity.

Figure 3 shows the calculated isotherms and the temperature profiles at four downstream locations for the nontapered case for Amioca,  $n = 0.3$ . The parameter  $G$ , which represents the relative importance of viscous dissipation compared to heat conduction in the  $y$ -direction is quite high and gives rise to fluid temperatures that are higher than the barrel temperature (i.e.  $\theta > 1$ ). Consequently, heat transfer occurs from the fluid to the barrel, which may, therefore, have to be cooled in the sections close to the die to ensure that it is maintained at a particular temperature. For smaller values of  $G$ , the more common situation of heat transfer from the barrel to the fluid arises. The temperature gradient,  $\partial T / \partial y$ , at the barrel is also higher than that at the screw root because the heat generated by viscous dissipation has to be conducted away from the barrel to the ambient, whereas the screw root is adiabatic. Material temperatures are typically seen to be higher than imposed barrel temperatures, and this has important implications in extruder design.

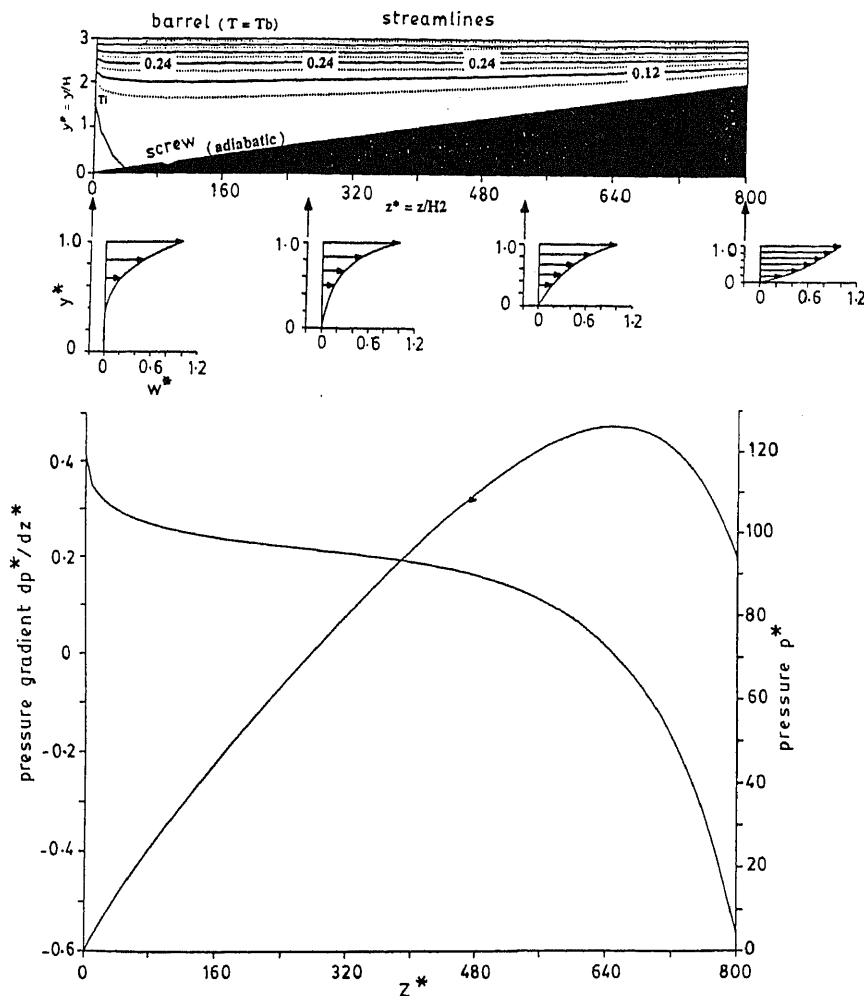
Figure 4 shows the moisture concentration contours. The effect of the sink term



**Figure 3.** Isotherms and temperature profiles (a) at four downstream locations for Amioca with  $n=0.3$ ,  $\phi=16.54^\circ$ ,  $q_v=0.25$ ,  $Pe=7050$ ,  $G=0.005$ ,  $\beta=1.134$ ,  $\beta_1=10.0$ ,  $b_m=1.0$ ,  $S=-1000$ ,  $Le=0.001$ ,  $\theta_{gel}=0.5$ , no taper. Also shown is the velocity field (b).



**Figure 4.** Lines of constant moisture concentration  $c^*$ , along with the corresponding profiles, at four downstream locations for Amioca for the conditions given in figure 3.



**Figure 5.** Streamlines and velocity profiles at four downstream locations, and pressure and pressure gradient variation with  $z^*$  for Hylon 7 in a tapered channel with  $n=0.5$ ,  $q_w=0.2$ ,  $Pe=3000$ ,  $G=1.0$ ,  $\beta=1.134$ ,  $\beta_1=5.0$ ,  $b_m=0.0$ , and  $\psi=0.1432^\circ$ .

$S$ , due to the reaction, is manifested in the form of a decrease in the moisture concentration due to bonding of water for gel formation as the material reaches the gelatinization temperature  $\theta_{gel}$ . This loss of moisture occurs at a rate specified by the sink term  $S$  and is obviously more rapid for larger  $S$ . As the sink term is increased in magnitude from  $S=0$  (i.e. no reaction), the sharp decrease in the moisture concentration occurs earlier along the screw channel as expected, and also occurs first at the screw root. The two variables that control the magnitude of viscosity – that is, the temperature and the moisture concentration – act in opposing directions as the material is heated and thus loses moisture as a result of gelatinization. The temperature profiles indicate that the screw root gets hotter than the barrel wall for the set of conditions shown in figure 4, and this leads directly to gelatinization near the screw first. However, a reaction rate that is a function of both temperature and shear is a more realistic representation of the cooking process. In such cases, the

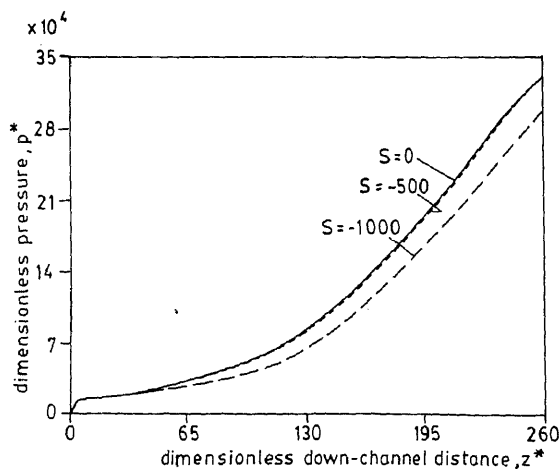
gelatinization contours may be quite different from the ones shown here. Effort is currently underway to incorporate thermal and shear effects into the model.

The results of the simulation for a typical taper angle of  $0.1432^\circ$  are shown in figure 5. This taper angle corresponds to the Brabender single-screw extruder. The screw root is continuously tapered from the feed section to the die. The normalization for all lengths is based on the final channel depth ( $H_2$ ) except for the vertical screw channel distance,  $y$ , which is scaled with the local depth ( $H$ ). In figure 5, the taper section is enlarged for clarity.

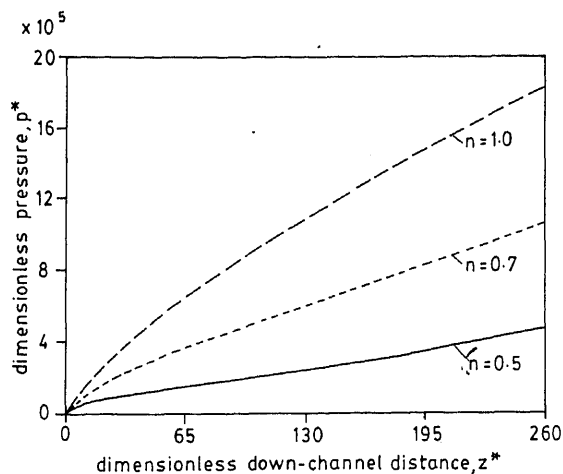
Consider the case of the isothermal flow of a Newtonian fluid in a rectangular, tapered channel. Depending on the imposed pressures at either end of the channel, there could be a maximum in the pressure profile somewhere in between. This can be shown analytically. There is pure drag flow at the location of this maximum. For non-Newtonian flow, however, it is not clear whether this maximum necessarily should occur inside the extruder or outside it.

For the nonisothermal case, as the material flows in a tapered channel, it gets squeezed in the decreasing gap. However, the net flow rate has to be maintained at the same value at any channel cross section. As shown in figure 5, the spacing of the streamlines is closer toward the die. The velocity profiles at the four downstream locations show markedly different behaviour for tapered channels compared to the case with no taper. It may be noted that the initially curved velocity profile becomes linear (corresponding to drag flow) as the section becomes shallower. At the die section, the profile shows velocities in the channel that are larger than the barrel velocity (or the screw tip speed, in the screw-moving formulation), again demonstrating the squeezing effect. The temperature profile becomes almost completely uniform across the channel depth in the regions close to the die for a tapered channel. This temperature is also equal to the imposed barrel temperature. Compared to the case with no taper, this shows that the narrower gap results in better thermal mixing and uniformity. These results can be used advantageously in extruder design for situations in which a specific pressure level needs to be developed while maintaining good product mixing facilitated by the squeezing effect of the tapered channel.

Figure 6 shows the effect of varying the strength of the sink on the pressure rise along the channel. As compared to the case of  $S = 0$ , the dimensionless pressure at the die is seen to be lower for higher sink strengths. The variation of the downstream



**Figure 6.** Effect of the strength of the moisture sink  $S$  on the variation of the dimensionless pressure  $p^*$  along the screw channel length  $z^*$  for the conditions given in figure 3.

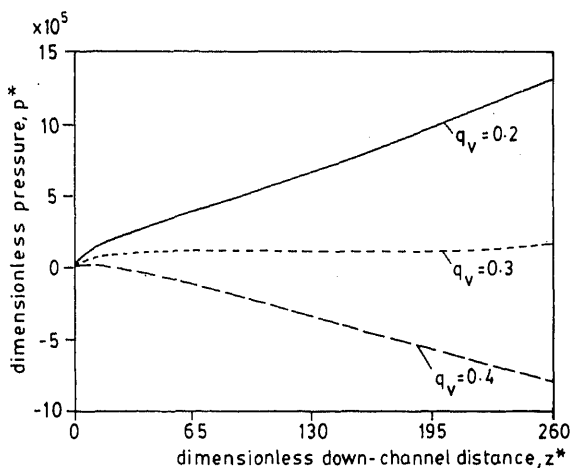


**Figure 7.** Effect of the power law index  $n$  on the variation of dimensionless pressure  $p^*$  along the screw channel length  $z^*$  for  $q_v = 0.3$ ,  $G = 0.001$ , and other conditions as in figure 3.

pressure gradient (not shown here) with the channel length shows that there is a kink in the profile at the point where the gelatinization first occurs. However, this rise in the value of  $dp/dz$  is not strong enough to counter the temperature effect, which tends to lower the viscosity and, thus, makes the material flow more easily.

Figure 7 shows the effect on the pressure obtained in the extruder channel of changing the power law index  $n$  of the material. Starches can be typically characterized by power law indices of less than 0.5. The curves are for  $G = 0.001$ ,  $q_v = 0.3$ , and  $S = -1000$ . Thus, only  $n$  is varied, keeping the rest of the terms in (4) the same. This implies a higher viscosity for higher  $n$  at a given shear rate. These results indicate the effect of  $n$ , though no actual physical circumstance is simulated here. The pressure rises from the hopper to the die along the screw helix, as expected. The pressure gradient is higher for increasing  $n$ . For non-Newtonian fluids,  $n < 1$ , and the viscosity decreases with an increase in shear rate. The Newtonian fluid,  $n = 1$ , therefore, gives rise to larger viscous drag, which in turn means that larger pressure gradients are required to overcome it.

The effect of the throughput  $q_v$  on the pressure development at the die is investigated next. As shown in figure 8, a smaller value of  $q_v$ , which corresponds to greater



**Figure 8.** Effect of the throughput  $q_v$  on the variation of dimensionless pressure  $p^*$  along the screw channel length  $z^*$  for  $n = 0.5$ ,  $G = 0.0$ , and other conditions as in figure 3.

restriction to the flow at the die, results in a larger pressure rise as compared to the case with  $q_v = 0.4$ , which is close to the open die situation for the isothermal flow of Newtonian fluids,  $q_v = 0.5$ . For the particular set of conditions here, a larger value of  $q_v$  results in pressure loss from the hopper to the die. This is because the balance between the pressure buildup as the material flows through the extruder channel, and the easing of pressure due to the decrease of viscosity is not maintained. If the flow rate is increased, the pressure at the die decreases further. Therefore, there is a limiting value of the throughput  $q_v$  that can be fed through an extruder beyond which pressure development and, correspondingly, the development of desirable material quality, are affected considerably. This limiting value depends on the particular conditions governing the flow: for example, the power law index  $n$  and the viscosity coefficients  $\beta$ ,  $\beta_1$ , and  $b_m$ . For isothermal, Newtonian flow, a simple theoretical analysis shows that  $q_v = 0.5$  corresponds to the throughput in the extruder in the absence of a die. For the nonisothermal, polymeric flow considered here, this value for the no-die situation is difficult to estimate, but is typically around 0.3.

## 5. Conclusions

A numerical simulation of the transport phenomena occurring in the flow of food materials through a single-screw extruder has been carried out. For significant viscous dissipation within the material, the material temperature rises above the imposed barrel temperature by as much as 100%. The flow field, however, is not strongly affected by this change. The moisture concentration contours provide a quantitative estimate of the extent of cooking that the food material undergoes. Moisture removal and bonding due to gelatinization of starch take place earlier and first at the screw root for higher reaction rate. Gelatinization increases the viscosity significantly, which in turn causes larger viscous heating and, consequently, cooking of the food material. Pressure development at the end of the extruder is significantly affected by the throughput. For throughputs larger than a limiting value, typically 0.3, the pressure can actually decrease from the hopper to the die, which means that the material has to be pumped through the extruder. Tapering of the screw channel is an important means of controlling the pressure development at the die.

This is publication No. F-10544-4-91 of the New Jersey Agricultural Experiment Station supported by State Funds and the Center for Advanced Food Technology (CAFT). The Center for Advanced Food Technology is a New Jersey Commission on Science and Technology Center. This work was also supported in part by the US Army Research Office. The authors would like to thank Dr M V Karwe and Professor V Sernas for discussions throughout this work.

## List of symbols

$b$	temperature coefficient of viscosity, (4);
$b_m$	$= B_m c_{mi}$ ;
$B_m$	moisture coefficient for viscosity;
$c_m$	moisture concentration;
$c_{mi}$	initial moisture concentration;



$c^*$	dimensionless $c_m$ ;
$C$	specific heat of the fluid;
$D$	mass diffusivity of moisture into the bulk material;
$D_b$	barrel diameter;
$G$	Griffith number, $G = \bar{\mu} V_{bz}^2 / k (T_b - T_i)$ ;
$H$	height of the screw channel, as a function of $z$ ;
$H_1$	initial screw channel height;
$H_2$	final screw channel height;
$k$	thermal conductivity of the fluid;
$L$	axial screw length;
$Le$	Lewis number, representing the relative magnitude of mass diffusivity to thermal diffusivity;
$n$	power law index, (4);
$N$	screw speed (rpm);
$p$	pressure;
$\bar{p}$	reference pressure, $\bar{p} = \bar{\mu} V_{bz} / H_2$ ;
$p^*$	dimensionless pressure;
$Pe$	Peclet number, $Pe = V_{bz} H_2 / \alpha$ ;
$q_v$	dimensionless volumetric flow rate (throughput);
$Q$	total volumetric flow rate;
$S$	dimensionless sink term for moisture;
$s'$	rate of reaction;
$t$	time;
$\bar{t}$	average residence time;
$T$	temperature;
$T_b$	barrel temperature;
$T_i$	inlet temperature;
$u$	velocity component in $x$ -direction;
$v$	velocity component in $y$ -direction;
$V_b$	tangential barrel velocity, $V_b = \pi D_b N / 60$ ;
$V_{bx}$	component of $V_b$ along $x$ , $V_{bx} = V_b \sin(\phi)$ ;
$V_{bz}$	component of $V_b$ along $z$ , $V_{bz} = V_b \cos(\phi)$ ;
$w$	velocity component in $z$ -direction;
$W$	width of the screw channel;
$x$	coordinate axis normal to screw flights;
$y$	coordinate axis normal to the screw root;
$z$	coordinate axis along the screw channel;
$\alpha$	thermal diffusivity, $\alpha = k / \rho C$ ;
$\beta$	dimensionless $T_b$ , $\beta = T_b / T_i$ ;
$\beta_1$	dimensionless $b$ , $\beta_1 = b / T_i$ ;
$\phi$	screw helix angle;
$\dot{\gamma}$	strain rate;
$\dot{\gamma}_0$	reference strain rate, (4);
$\dot{\gamma}^*$	dimensionless strain rate;
$\mu_0$	reference viscosity, (4);
$\bar{\mu}$	average viscosity, $\bar{\mu} = \mu_0 [V_{bz} / H_2 \dot{\gamma}_0]^{n-1}$ ;
$\psi$	taper angle for a tapered screw channel;
$\theta$	dimensionless temperature, $\theta = (T - T_i) / (T_b - T_i)$
$\rho$	density of the fluid;
$\tau$	shear stress.

*Superscripts*

\* dimensionless quantity.

**References**

- Elbirli B, Lindt J T 1984 A note on numerical treatment of the thermally developing flow in screw extruders. *Polym. Eng. Sci.* 24: 482-487
- Fenner R T 1977 Developments in the analysis of steady screw extrusion of polymers. *Polymer* 18: 617-622
- Fenner R T 1979 *Principles of polymer processing* (New York: Chemical Publishing)
- Harper J 1980 *Food extrusion* (Boca Raton, FL: CRC Press)
- Jaluria Y 1988 *Computer methods for engineering* (Needham Heights, MA: Allyn & Bacon)
- Jaluria Y, Torrance K E 1986 *Computational heat transfer* (New York: Hemisphere)
- Karwe M V, Jaluria Y 1990 Numerical simulation of fluid flow and heat transfer in a single screw extruder for non-Newtonian fluids. *Numer. Heat Transfer* A17: 167-190
- Kays W M, Crawford M E 1980 *Convective heat and mass transfer* 2nd edn (New York: McGraw-Hill)
- Kokini J L 1987 Physical forces in food systems. Research accomplishment report, Center for Advanced Food Technology, Rutgers University, New Brunswick, New Jersey
- Schlichting H 1979 *Boundary layer theory* (New York: McGraw-Hill)
- Tadmor Z, Gogos C 1979 *Principles of polymer processing* (New York: Wiley)
- Wang S S, Chiang C C, Yeh A I, Zhao B, Kim I H 1989 Kinetics of phase transition of waxy corn starch at extrusion temperatures and moisture contents. *J. Food Sci.* 54: 1298-1301

# A novel enthalpy formulation for multidimensional solidification and melting of a pure substance

A W DATE

Mechanical Engineering Department, Indian Institute of Technology,  
Powai, Bombay 400 076, India

**Abstract.** This paper presents a new finite-difference formulation of the multidimensional phase change problems involving unique phase change temperature. The solutions obtained with this formulation show that the problem of “waviness” of the temperature histories encountered with the conventional enthalpy formulation is now removed. The formulation derived provides a simple method for “local” tracking of the interface using the enthalpy variable in a novel way. During the solution of the finite-difference equations, the present formulation obviates the need for “book-keeping” of the phase-change nodes, and hence allows solution of the equations by tridiagonal matrix algorithm. It is argued that the benefits of enthalpy formulation can be extended to phase-change problems involving convection by solving the equations of motion on non-staggered grid.

**Keywords.** Enthalpy formulation; multidimensional phase-change problems; finite-difference method.

## 1. Introduction

### 1.1 *The problem considered*

There is hardly any engineering product that, during its manufacture, does not undergo a process of solidification and melting at some stage. Engineering processes such as casting, welding, surface alloying, dip forming, crystallisation etc. involve phase-change. The process of freezing and thawing is of interest in preservation of foods. The phenomenon of “permafrost” is concerned with changes in load-bearing capacity of soils in very cold environments. The principle of latent heat transfer is used in the development of compact thermal energy storage devices that enable storage and retrieval of energy at nearly constant temperature.

The phenomenon of solidification or melting is brought about by a process of latent heat ( $\lambda$ ) transfer at the interface between the solid and the liquid phases. For a pure substance, throughout this process, the temperature  $T_m$  of the interface remains constant. Both  $\lambda$  and  $T_m$  are properties of a pure substance. Within each of the single

phase, the heat transfer is essentially governed by the process of unsteady conduction, although under certain circumstances, convection may also occur in the liquid phase under the action of body (e.g. buoyancy) or surface (e.g. surface tension) forces.

There are two approaches to solving the phase change problems:

- (i) the variable domain formulation;
- (ii) the fixed domain formulation.

In the first approach, which has several variants, two energy equations in their conventional form are solved in the solid and the liquid phases with temperatures  $T_s$  and  $T_l$  as the dependent variables respectively. In addition to the domain boundary and initial conditions, the following two conditions are imposed at the interface:

$$T_s = T_l = T_m, \quad (1)$$

$$K \frac{\partial T_s}{\partial \mathbf{n}} \Big|_i - K \frac{\partial T_l}{\partial \mathbf{n}} \Big|_i = \rho \lambda v_i, \quad (2)$$

where  $\mathbf{n}$  is a vector normal to the interface and  $v_i$  is the instantaneous velocity of the interface in the direction of the normal.

Since the method requires continuous tracking of the interface, the physical coordinates are usually normalised with respect to this location. The governing energy equations are then freshly derived in the transformed coordinate system in which the interface is immobilised, although in the physical coordinate system the volumes occupied by each of the single phases change with time (and, hence, the designation of variable domain formulation). In multidimensional problems, this transformation involves laborious algebra that gives rise to pseudo-convection terms and cross-derivatives (see, for example, Saitoh 1978), the former may even give rise to numerical instabilities under certain circumstances. The finite difference implementation of the variable domain formulation can be carried out through several variants that are described by Basu & Date (1988); not all of which are amenable to easy extension to multidimensional problems.

This paper is concerned with the second type of formulation, namely the fixed domain or the enthalpy formulation, which treats the total enthalpy  $H$ , rather than the temperature  $T$ , as the *main* dependent variable in the energy equation. Now since enthalpy is a conserved property, the energy equations for both phases can be written in terms of a single equation, viz:

$$\frac{\partial}{\partial t}(\rho H) + \text{div}(\rho \mathbf{u} H) = \text{div}(K \text{ grad } T), \quad (3)$$

where  $\mathbf{u}$  is the velocity vector which may be finite in the liquid phase.

Incidentally, it can be shown that (3) already satisfies the interface flux condition (2). As such, the equation applies to the entire domain of interest and the interface need not be tracked *during the differential* formulation of the phase-change problem. Hence, this formulation is also known as the fixed domain formulation. Note, however, that (3) contains two dependent variables,  $H$  and  $T$ , and a set of auxiliary relations (also known as equations of state) between them must be specified.

In several applications, the domain boundaries are often of complex shape. Also because of the asymmetries of the boundary shapes, and thermal boundary conditions, and because of the presence of convection, the interface, during its evolution, can

assume complex shapes. It has become increasingly apparent that such complexities (Wilson *et al* 1978; Sparrow *et al* 1988, pp. 747–86) can best be handled by the enthalpy formulation; rather than the variable domain formulation. This is particularly so if it is of interest to develop generalised computer codes for the general problem of solidification and melting.

Computationally speaking, it is relatively simple to implement the enthalpy formulation via discretised equations when the substance is impure. For, in this case, the latent heat transfer takes place over a range of temperatures that demarcate what is known as the “mushy” region. The physical and transport properties of this region must however be known, or modelled. For a pure substance, however, the phase-change takes place at a unique temperature and it was shown by Voller *et al* (1979), that unless special procedures are adopted, the predicted temperature and heat flux histories, as well as the interface movement, are unrealistic. Recently, Voller (1990) has reviewed several implicit procedures using the enthalpy formulation. Date (1991, 1992) has also reviewed some of the earlier methods and identified their shortcomings.

Recently, Date (1992) has presented an enthalpy formulation that eliminates the problem of prediction of unrealistic temperature histories, allows use of an efficient line-by-line numerical integration algorithm and is applicable to multidimensional problems. The purpose of this paper is to present Date’s formulation giving further details than those given in Date (1992).

## 1.2 Outline of the paper

The paper is divided into five sections. Section 2 describes Date’s (1992) formulation in one and two dimensions, § 3 deals with phase-change problems involving convection. Here it is argued that since the interface can assume arbitrary shape, solution of equations of motion on non-staggered grids offers considerably more computational convenience than the use of staggered grids. A few illustrative solutions to the phase-change problems are presented in § 4. Finally conclusions are reported in § 5.

## 2. Enthalpy formulation of Date (1992)

### 2.1 One-dimensional problems

For the purposes of discussion, we consider one-dimensional heat transfer without bulk convection; further assuming uniform properties. Equation (3) can then be written as:

$$\rho \frac{\partial H}{\partial t} = k \frac{\partial^2 T}{\partial X^2}. \quad (4)$$

In order to solve the above equation,  $H$  must be replaced by  $T$  or vice versa. This is done via the equations of state which provide the  $H$ – $T$  relationship. Figure 1 shows this relationship which is only piece-wise continuous. Mathematically, the relationship can be written in two ways as shown in table 1. Most of the previous authors have preferred the  $H = f(T)$  relationship; we prefer the  $T = f(H)$  relationship following Shamsunder & Sparrow (1975).

It is convenient to define the following dimensionless variables:

$$\phi = (H - H_s)/\lambda, \quad (5)$$

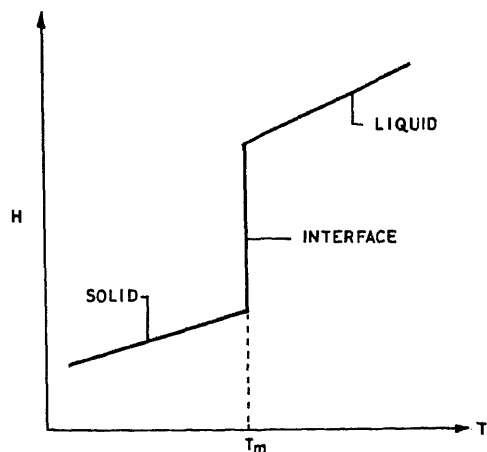


Figure 1.  $H$ - $T$  relationship for a pure substance.

$$\theta = C_p(T - T_m)/\lambda, \quad (6)$$

$$X = x/L, \quad (7)$$

$$\tau = \alpha t/L^2. \quad (8)$$

In the new variables, (4) reads as:

$$\frac{\partial \phi}{\partial \tau} = \frac{\partial^2 \theta}{\partial X^2}; \quad (9)$$

and the  $T = f(H)$  relationship reads as:

$$\theta = \phi; \quad \theta \leq 0 \quad (\text{solid}), \quad (10)$$

$$\theta = 0; \quad 0 \leq \phi \leq 1 \quad (\text{interface}), \quad (11)$$

$$\theta = \phi - 1; \quad \phi \geq 1 \quad (\text{liquid}). \quad (12)$$

Relationships (10) to (12) can be generalised as:

$$\theta = \phi + \phi', \quad (13)$$

Table 1.  $H$ - $T$  relationship.

	$H = f(T)$	$T = f(H)$
Solid	$H = C_p T; T \leq T_m$	$T = H/C_p; H \leq H_s$
Interface	$H = C_p T_m + H_{ps}(t); T = T_m$	$T = T_m; H_s \leq H \leq H_1$
Liquid	$H = C_p T + \lambda; T \geq T_m$	$T = (H - \lambda)/C_p; H \geq H_1$
	$\int_t^{t+t'} \frac{dH_{ps}}{dt} dt = H_1 - H_s = \lambda$ <p>where <math>t'</math> = time for latent heat transfer</p>	

where

$$\phi' = 0.5\{|1 - \phi| - |\phi| - 1\}. \quad (14)$$

Equation (14) ensures that  $\phi' = 0$  in solid,  $\phi' = -\phi$  at the interface and  $\phi' = -1$  in liquid.

Now (9) can be finite-differenced via control-volume analysis as:

$$\phi_j^n = \phi_j^o = S(\theta_{j+1}^n - 2\theta_j^n + \theta_{j-1}^n); \quad (15)$$

where

$$S = \Delta\tau/\Delta X^2; \quad (16)$$

and  $j$  identifies the grid node, whereas superscripts  $n$  and  $o$  identify the new and the old values. In deriving (15), uniform grid spacing  $\Delta X$  is assumed.

Equation (13) can now be used to replace  $\theta$ 's in (15), so that:

$$\phi_j^n(1 + 2S) = S(\phi_{j+1}^n + \phi_{j-1}^n) + S(\phi_{j+1}^{n-1} - 2\phi_j^{n-1} + \phi_{j-1}^{n-1}) + \phi_j^o. \quad (17)$$

In the above equation,  $\phi'$  values lag behind the  $\phi$  values by one iteration. As such the  $\phi'$  terms (along with  $\phi_j^o$ ) can be treated as sources. Equation (17) is then unconditionally stable according to the Scarborough criterion, and can be solved by the point-by-point Gauss-Seidel scheme or by the line-by-line Tridiagonal Matrix algorithm (TDMA). Further, it is not necessary to carry out "book-keeping" of the nodes which are in solid, liquid and phase-change states.

The formulation is thus an improvement over that of Shamsunder & Sparrow (1975) who did not generalise the  $T = f(H)$  relationship in the manner of (13) and (14) and therefore had to resort to node "book-keeping" which necessitates the use of only point-by-point integration procedure. Basu & Date (1987) have however shown that application of TDMA results in much faster convergence than the point-by-point procedure particularly when fine mesh size is used.

This formulation however suffers from one drawback. It will be recognised that  $0 < \phi < 1$  at the phase-change node. As such, throughout the period of transition of the interface through the control-volume surrounding the phase-change node  $\phi' = -\phi$ , and therefore the nodal value of the phase-change node remains stationary at  $\theta = \phi + \phi' = 0$ . As a result, the predicted temperature histories demonstrate a step-like or a wavy pattern (Shamsunder 1978, pp. 165-83; Voller *et al* 1979).

One way in which this peculiar behaviour can be eliminated is to use a fine mesh (or small  $\Delta X$ ); so that the time period over which  $\theta$  remains stationary at zero is minimised, and the essentially wavy solutions appear to be smooth and accurate. This measure was adopted by Shamsunder (1978, pp. 165-83), but it exacts penalty in computer time which can partly be compensated by the application of TDMA, as described above.

To prevent  $\theta$  from remaining stationary at zero at the phase change node, it is necessary to rewrite the generalised  $T = f(H)$  relationship as:

$$\theta = \phi + \phi'', \quad (18)$$

where

$$\phi'' = \phi' + \theta_{pc}. \quad (19)$$

Here  $\theta_{pc}$  denotes the nodal value of the temperature at the phase-change node; it equals zero at the single phase nodes. With the replacement suggested by (18), (15)

can now be written as:

$$\phi_j^n(1 + 2S) = S(\phi_{j+1}^n + \phi_{j-1}^n) + S(\phi_{j+1}^{n'} - 2\phi_j^{n'} + \phi_{j-1}^{n'}) + \phi_j^o. \quad (20)$$

The above equation is same as (17), except that  $\phi'$  is replaced by  $\phi''$ . The  $\phi''$  values again lag behind the  $\phi$  values by one iteration. Further, however, it becomes necessary to determine the value of  $\theta_{pc}$  at the phase-change node.

**2.1a Determination of  $\theta_{pc}$ :** Consider figure 2, where the phase change node  $j$  is shown along with nodes  $j-1$  and  $j+1$  which are in single phase. At the time instant considered, let the interface be located at a distance  $\Delta X_i$  to the east of the nodal position  $X_j$ . Now since the value of  $\theta$  at the interface is zero, one may linearly interpolate  $\theta_{pc,j}$  to read as

$$\theta_{pc,j} = \left[ \frac{\Delta X_i}{\Delta X_i + \Delta X} \right] \theta_{j-1}. \quad (21)$$

Similarly, if the interface is to the west of node  $j$  then:

$$\theta_{pc,j} = \left[ \frac{|\Delta X_i|}{|\Delta X_i| + \Delta X} \right] \theta_{j+1}, \quad (22)$$

where

$$\begin{aligned} \Delta X_i &= X_i - X_j \\ &= (0.5 - \phi_j)\Delta x \\ &= (0.5 + \phi_j'')\Delta X. \end{aligned} \quad (23)$$

Equations (21) and (22) can now be generalised as:

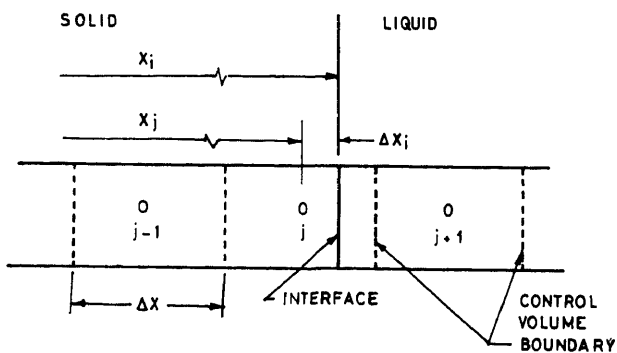
$$\theta_{pc,j} = 0.5[(A + |A|)\theta_{j-1} - (A - |A|)\theta_{j+1}]F, \quad (24)$$

where

$$A = \frac{0.5 + \phi_j''}{1 + |0.5 + \phi_j''|}; \quad (25)$$

and

$$F = -\frac{(1 + \phi_j')\phi_j'}{(1 - \phi_j)\phi_j}. \quad (26)$$



**Figure 2.** Typical phase change node  $j$ .



- 2.1b *Some pertinent comments:* (i) At the phase change node,  $\phi_j \Delta X$  represents the volume occupied by the liquid.
- (ii) The location of the interface is identified with  $\theta = \phi + \phi'' = 0$ . As such,  $\Delta X_i = 0.5 \Delta X - \phi_j \Delta X = (0.5 + \phi_j'') \Delta X$ , as shown by (23).
- (iii) Equation (24) ensures that when  $\Delta X_i$  is positive, (21) is used to determine the nodal value of temperature at the phase-change node. When  $\Delta X_i$  is negative, (22) is operational.
- (iv) The factor  $F$  given by (26) equals zero at the single phase nodes, but it equals one at the phase-change node [see (14)]. This ensures that  $\theta_{pc} = 0$  at single phase node but is finite at the phase change node.
- (v) Since  $\theta_{pc,j} = 0$  at the single-phase nodes  $\phi_j'' = \phi_j'$  at these nodes, that is  $\phi_j'' = 0$  in solid and  $\phi_j'' = -1$  in liquid. As such the location of  $X_i$  (or the volume occupied by the solid) can be calculated from:

$$X_i = \sum_{j=1}^N (1 + \phi_j'') \Delta X, \quad (27)$$

where  $N$  denotes the total number of nodes in the domain.

- (vi) Both  $\phi'$  and  $\theta_{pc}$  are calculated in such a way that it is not necessary to carry out "book-keeping" of the single-phase and phase-change nodes, and hence (20) can be solved by TDMA.

2.1c *The solution procedure:* The following steps are incorporated in the solution procedure:

- (1) Specify initial values of temperature  $\theta_{in}$  at all nodes.
- (2) Hence, evaluate  $\phi_{in}$  and  $\phi_{in}'$  at all nodes using (10) or (12) as appropriate, and (14). Set  $\phi_{in}'' = \phi_{in}'$ .
- (3) Perform one iteration of (20) for an arbitrarily chosen time step  $\Delta \tau$  to yield new values of  $\phi$ .
- (4) Evaluate  $\phi'$  from (14) and  $\theta_{pc}$  from (24) at all nodes using the just calculated values of  $\phi$ . Hence form  $\phi'' = \phi' + \theta_{pc}$  at all nodes.
- (5) Return to step (3) until  $\phi$  has converged between successive iterations.
- (6) Calculate  $\theta = \phi + \phi''$  at all nodes and evaluate  $X_i$  from (27) if desired.
- (7) Set  $\phi^o = \phi''$ , and return to step 3 to calculate the next time step.

The above strongly implicit procedure is stable for any value of the chosen time step  $\Delta \tau$ , and grid spacing  $\Delta X$ . Usually coarse grids (say 5 to 7 nodes) suffice. However, for high Stefan numbers, that is for high velocity of interface movement, finer grids (11 to 13 nodes) may be required to achieve good accuracy (Date 1992).

## 2.2 Two-dimensional problems

The dimensionless governing equation for such problems is given by

and the finite-difference analogue is given by:

$$\begin{aligned}\phi_{k,j}^n(1 + 2S_x + 2S_y) = & S_x(\phi_{k+1,j}^n + \phi_{k-1,j}^n) \\ & + S_y(\phi_{k,j+1}^n + \phi_{k,j-1}^n) \\ & + S_x(\phi_{k+1,j}^{''n} - 2\phi_{k,j}^{''n} + \phi_{k-1,j}^{''n}) \\ & + S_y(\phi_{k,j+1}^{''n} - 2\phi_{k,j}^{''n} + \phi_{k,j-1}^{''n}) \\ & + \phi_{k,j}^0,\end{aligned}\quad (29)$$

where

$$\begin{aligned}S_x &= \Delta\tau/\Delta X^2, \\ S_y &= \Delta\tau/\Delta Y^2.\end{aligned}\quad (30)$$

**2.2a Determination of  $\theta_{pc}$ :** In order to evaluate  $\phi_{k,j}^{''n}$ , it becomes necessary to estimate  $\theta_{pc,kj}$  at all phase-change nodes. Note that, in a multi-dimensional problem, several nodes may undergo phase change simultaneously. The procedure adopted for evaluation of  $\theta_{pc,kj}$  is as follows:

$$\theta_{pc,kj} = 0.5(\theta_{pcx} + \theta_{pcy})F, \quad (31)$$

where

$$F = -\{(1 + \phi'_{kj})\phi'_{kj}\}/\{(1 - \phi_{kj})\phi_{kj}\}. \quad (32)$$

Now the  $X$  and  $Y$  components of  $\theta_{pc}$  are evaluated as follows:

$$\theta_{pcx} = 0.5\left[\left(\frac{Bx}{|Bx|} + 1\right)\theta_{pcx_1} - \left(\frac{Bx}{|Bx|} - 1\right)\theta_{pcx_2}\right], \quad (33)$$

where

$$Bx = |Ax| - 0.5; \quad (33a)$$

$$Ax = \Delta X_i / \{|\Delta X_i| + \Delta X\}, \quad (33b)$$

$$\theta_{pcx_1} = 0.5(\theta_{k-1,j} + \theta_{k+1,j}); \quad (33c)$$

and

$$\theta_{pcx_2} = 0.5[(Ax + |Ax|)\theta_{k-1,j} - (Ax - |Ax|)\theta_{k+1,j}]. \quad (33d)$$

Similarly

$$\theta_{pcy} = 0.5\left[\left(\frac{By}{|By|} + 1\right)\theta_{pcy_1} - \left(\frac{By}{|By|} - 1\right)\theta_{pcy_2}\right], \quad (34)$$

where

$$By = |Ay| - 0.5, \quad (34a)$$

$$Ay = \Delta Y_i / \{|\Delta Y_i| + \Delta Y\}; \quad (34b)$$

$$\theta_{pcy_1} = 0.5(\theta_{k,j-1} + \theta_{k,j+1}), \quad (34c)$$

and

$$\theta_{pcy_2} = 0.5[(Ay + |Ay|)\theta_{k,j-1} - (Ay - |Ay|)\theta_{k,j+1}]. \quad (34d)$$

Here  $\Delta X_i$  and  $\Delta Y_i$  represent the location of the interface along  $X$  and  $Y$  axes respectively (see figure 3). Note that in a general phase-change problem, the interface may occupy the phase-change control volume in any of the six ways shown in figure 3.

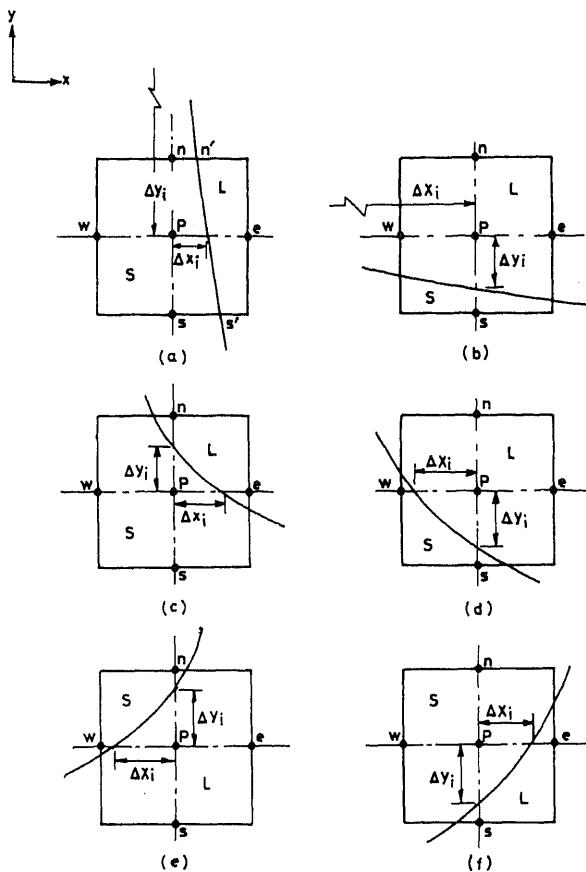


Figure 3. Different interface positions.

Equations (33) and (34) show that when  $|\Delta X_i| \geq \Delta X$  or  $|\Delta Y_i| \geq \Delta Y$  then  $\theta_{pcx} = \theta_{pcx1}$  and  $\theta_{pcy} = \theta_{pcy1}$ ; when  $|\Delta X_i| \leq \Delta X$  and  $|\Delta Y_i| \leq \Delta Y$ ,  $\theta_{pcx} = \theta_{pcx2}$  and  $\theta_{pcy} = \theta_{pcy2}$ . Thus, one-sided interpolation for  $\theta_{pc}$  is carried out only when the interface lies between the phase-change node and its immediate neighbouring node.

**2.2b Determination of  $\Delta X_i$  and  $\Delta Y_i$ :** It is clear from figure 3 that the expressions for  $\Delta X_i$  and  $\Delta Y_i$  will depend on the type of intersection of the interface with the control volume. Further, it must be possible to determine the type of intersection itself. Both these are accomplished by using  $\phi'$  and  $\phi''$  variables. Here expressions for Type (a) and Type (c) intersections will be devised; for other types, the appropriate expressions are given in the appendix A.

Type (a) intersection is thus identified by:

$$\begin{aligned}\phi'_{k+1,j} &= -1; \\ \phi'_{k-1,j} &= 0.\end{aligned}\tag{35}$$

Further, area  $nn'ss' = (0.5 + \phi''_{kj})\Delta X\Delta Y$ , or

$$\Delta X_i = (0.5 + \phi''_{kj})\Delta X.\tag{35a}$$

In order to determine  $\Delta Y_i$ , the procedure adopted is as follows:

$$\Delta Y_i = - \frac{\partial \Delta Y_i}{\partial X} \bigg|_{kj} \Delta X_i - \frac{\partial^2 \Delta Y_i}{\partial X^2} \bigg|_{kj} \frac{\Delta X_i^2}{2} + \text{neglected terms}$$

Now

$$\frac{\partial \Delta Y_i}{\partial X} = 1 / \left( \frac{\partial \Delta X_i}{\partial Y} \right) = -1 / \left( \frac{\partial \phi}{\partial Y} \bigg|_{kj} \right) \Delta X = 1 / \left( \frac{\partial \phi'}{\partial Y} \bigg|_{kj} \right) \Delta X,$$

and

$$\frac{\partial^2 \Delta Y_i}{\partial X^2} = - \left\{ \frac{\partial^2 \Delta X_i}{\partial X \partial Y} / \left( \frac{\partial \Delta X_i}{\partial Y} \right)^2 \right\} \bigg|_{kj} = - \left\{ \frac{\partial^2 \phi'}{\partial X \partial Y} \bigg|_{kj} / \left( \frac{\partial \phi'}{\partial Y} \bigg|_{kj} \right)^2 \Delta X \right\}$$

Thus,

$$\Delta Y_i = -C_1 + \frac{C_1^2 \Delta X}{2} \frac{\partial^2 \phi'}{\partial X \partial Y} \bigg|_{kj}, \quad (35b)$$

where

$$C_1 = (0.5 + \phi''_{kj}) / (\partial \phi' / \partial Y) \bigg|_{kj}.$$

The derivatives of  $\phi'$  are evaluated by central difference. Equation (35b) shows that when the interface is parallel to the  $Y$ -axis (as in a one-dimensional problem)  $\partial \phi' / \partial Y = 0$  and  $\Delta Y_i \rightarrow \infty$  as would be expected; and (35a) is the same as (23).

Type (c) intersection is identified by:

$$\phi'_{k-1,j} = \phi'_{k,j-1} = 0. \quad (36)$$

Further,

$$\Delta Y_i = (Y_i - Y_n) + 0.5 \Delta Y, \quad (36a)$$

where

$$Y_i - Y_n = - \frac{\partial \Delta Y_i}{\partial X} \bigg|_n \Delta X - \frac{\partial^2 \Delta Y_i}{\partial X^2} \bigg|_n \frac{\Delta X_{in}^2}{2} + \dots \quad (36b)$$

$$\Delta X_{in} = X_{in} - X_n = (0.5 + \phi''_n) \Delta X; \quad (36c)$$

$$\phi''_n = \phi''_{kj} + \phi''_{k,j+1} + \frac{1}{8} [\phi''_{k-1,j+1} + \phi''_{k-1,j} + \phi''_{k+1,j+1} + \phi''_{k+1,j}]. \quad (36d)$$

$$\frac{\partial \Delta Y_i}{\partial X} \bigg|_n = 1 / \left\{ \left( \frac{\partial \phi'}{\partial Y} \bigg|_n \right) \Delta X \right\}; \quad (36e)$$

and

$$\frac{\partial^2 \Delta Y_i}{\partial X^2} \bigg|_n = - \left\{ \frac{\partial^2 \phi'}{\partial X \partial Y} / \left( \frac{\partial \phi'}{\partial Y} \right) \Delta X \right\} \bigg|_n. \quad (36f)$$

Then

$$\begin{aligned} \Delta Y_i = & 0.5 \Delta Y - (0.5 + \phi''_n) / \left\{ \frac{\partial \phi'}{\partial Y} \bigg|_n \right\} \\ & + \left\{ (0.5 + \phi''_n)^2 \frac{\partial^2 \phi'}{\partial X \partial Y} \bigg|_n \Delta X \right\} / 2 \left( \frac{\partial \phi'}{\partial Y} \bigg|_n \right)^2. \end{aligned} \quad (36g)$$

Similarly, it can be shown that:

$$\Delta X_i = 0.5\Delta X - (0.5 + \phi_e'') \left/ \left\{ \frac{\partial \phi'}{\partial X} \right|_e \right\} + \left\{ (0.5 + \phi_e'')^2 \frac{\partial^2 \phi'}{\partial X \partial Y} \right|_e \Delta Y \right\} / 2 \left( \frac{\partial \phi'}{\partial X} \right|_e \right)^2, \quad (36h)$$

**2.2c The solution procedure:** Thus once  $\Delta X_i$  and  $\Delta Y_i$  are determined for each phase-change node,  $\theta_{pc}$  can be determined from (31) to (34). The overall solution procedure is given below:

- (1) Specify initial values of temperature  $\theta_{in}$  at all nodes.
- (2) Hence evaluate  $\phi_{in}$  and  $\phi'_{in}$  and set  $\phi''_{in} = \phi'_{in}$ .
- (3) Perform one iteration of (29) using double-sweep TDMA for an arbitrarily chosen time step  $\Delta \tau$  to yield new values of  $\phi$ . Calculate  $\phi'$  values at all nodes.
- (4) Identify the type of interface intersection from  $\phi'$  distribution. Hence calculate  $\Delta X_i$  and  $\Delta Y_i$  from appropriate expression.
- (5) Calculate  $\theta_{pc,kj}$  from  $\Delta X_i$  and  $\Delta Y_i$ .
- (6) Form  $\phi''_{kj} = \phi'_{kj} + \theta_{pc,kj}$ .
- (7) Return to step 3 until  $\phi$  has converged between successive iterations.
- (8) Calculate  $\theta = \phi + \phi''$  at all nodes.
- (9) Set  $\phi^o = \phi''$  and return to step 3 to calculate at the new time step.

### 2.3 Closure

In this section, an enthalpy formulation is developed, the unique feature of which is the generalisation of the  $T = f(H)$  relationship first in a continuum [(13), (14)], and then for a discretised domain [(18), (19)]. The latter is introduced to facilitate prediction of smooth temperature and heat flux histories without employing very fine mesh sizes.

The discretised version of the  $T = f(H)$  relationship required evaluation of  $\theta_{pc}$  at the phase-change nodes. This, in turn, requires location of the interface in terms of local coordinates  $\Delta X_i$  and  $\Delta Y_i$ . The  $\theta_{pc}$ ,  $\Delta X_i$  and  $\Delta Y_i$  are evaluated in terms of  $\phi'$  and  $\phi''$  which are functions of  $\phi$ . This method of locating the interface is unique in the sense that the interface may locally assume arbitrary shape and may intersect a given grid-line more than once. This contrasts with the method proposed by Patel (1968) (and which is used by Lazardis 1970 & Huang *et al* 1991) which attempts to predict interface coordinates  $X_i$  and  $Y_i$  relative to a fixed origin at all times by solving a differential equation. The present method for locating interface can also be extended to three-dimensional situations by employing appropriate Taylor-series expansion.

Further, the present method does not require "book-keeping" of the nodes in any of the operations and therefore enables use of line-by-line integration algorithm such as the TDMA.

### 3. Phase-change with convection

In several multidimensional phase-change problems, the operating or boundary conditions may be such that the liquid region experiences convection due to body or surface forces (Basu & Date 1988), requiring solution of the Navier-Stokes

equations to retrieve the velocity vector  $\mathbf{u}$  in (3). Usually such problems are solved by using the variable domain formulation (see, for example, Beckermann & Viskanta 1989). In such methods, since the liquid volume changes in size and shape with time, grids must be relaid at every time step with consequent requirement for interpolating the variables to the new positions assumed by the grid nodes. This is quite cumbersome. In the enthalpy formulation, since the grids remain fixed in space, such problems are avoided, although care is needed in effecting the boundary conditions on velocities at the phase-change nodes. This problem becomes particularly more complex when staggered grids are used for the velocity variables (see, for example, Gadgil & Gobin 1984 and Voller *et al* 1987) since the control-volumes surrounding the different velocity components occupy different amounts of solid and liquid, and which must be properly accounted. In order to fully realise the benefits of the fixed-grid enthalpy formulation, it is necessary to solve the Navier-Stokes equations on non-staggered grids.

During the last decade this has been achieved by employing in effect the concept of artificial compressibility (Rhie & Chow 1983) to eliminate the problem of checker-board prediction of pressure. Recently however, Date (1993, 1994) has pointed out some inelegant aspects of this method and its variants and suggested a new method that involves use of an effective pressure gradient to drive the nodal velocities. This method can be easily combined with the present enthalpy formulation since the liquid volume of any node is simply calculated as  $-\phi'_k \Delta X \Delta Y$  (in two dimensions). Also since  $\Delta X_i$  and  $\Delta Y_i$  are locally evaluated, they can be readily used to effect the no-slip boundary condition at the interface.

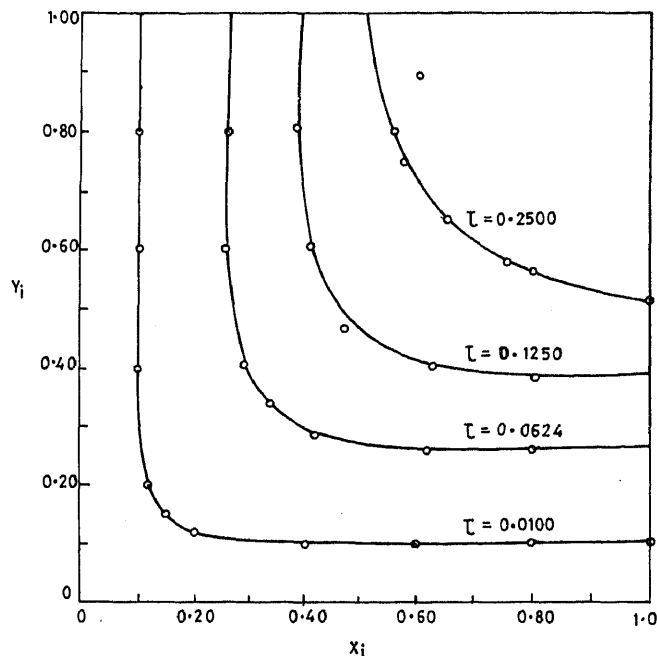
Complete details of present enthalpy formulation with convection are not given here for limitations of space, although these can be found in the dissertation by Pillay (1992) who has extended the formulation presented in §2 to include the effects of non-uniform grid-spacing, property difference of the two phases and convection driven by buoyancy.

#### 4. Some illustrative applications

The present enthalpy formulation has been applied to the solution of several one- and two-dimensional problems (without convection) involving temperature, heat flux and heat transfer coefficient boundary conditions and different initial conditions (Date 1992). Here, for the purposes of illustration, two 2-dimensional problems without convection and one problem with convection are considered.

*Problem 1:* This problem has been solved by Lazaridis (1970). Initially saturated liquid ( $\theta_{in} = 0$ ) is contained in a square domain of unit dimensions. At  $\tau = 0$ , the temperature at  $X = 0$  and  $Y = 0$  boundaries is lowered to  $\theta_o (= St) = -0.6405$  so that solidification commences instantly. The boundaries at  $X = 1$  and  $Y = 1$  are insulated.

Figure 4 shows the comparison of the predicted interface movement with that calculated by Lazaridis (1970). The present predictions are obtained with only five nodes in  $X$  and  $Y$  directions. The accuracy of predictions was checked by employing three time steps such that  $S_x = S_y = 0.0625, 0.125, 0.25$ . The computed results were found to be insensitive to this variation of time steps predicting the total solidification time at  $\tau = 0.59$ . For the largest time step the CPU time for complete solidification on CYBER 180/840 computer was 2.26 s, and that for the smallest time step was 9.02 s.



**Figure 4.** Evolution of interface - 2-D conduction phase change. [— Present ( $\Delta X = \Delta Y = 0.2$ ),  $\circ$  Lazaridis 1970.]

The special feature of this problem is that since  $\theta_0$  is constant along  $X = 0$  and  $Y = 0$  the predicted interface is symmetric about  $Y = X$  line and that at no time instant does the interface intersect a given grid line more than once.

**Problem 2:** In order to cause multiple intersections of the interface with a given grid-line at some instance of time, Lazaridis's problem was modified such that at  $X = 0$  and  $Y = 0$ ,  $\theta_0$  was assumed to vary linearly as:

$$\theta_0 = -(0.25 + 0.75 Z), \quad (37)$$

where

$$Z = X \text{ or } Y \text{ as appropriate.}$$

For this problem 8 nodes were used in  $X$  and  $Y$  directions. Figure 5 shows the interface movement as predicted by Pillay (1992). The total solidification time was predicted at  $\tau = 0.74$  and required CPU time of 10.4 s for  $S_x = S_y = 0.54$  and 13 s for  $S_x = S_y = 0.32$ . The computed results were again found to be independent of the time step. It is seen that at small times, the interface intersects a given grid line twice.

**Problem 3:** Beckermann & Viskanta (1989) performed experiments with melting of gallium. The test cell had inside dimensions of 4.76 cm height and width and 3.81 cm depth. Initially the gallium was at temperature  $T_c < T_m$ . At  $t = 0$ , the right vertical face is raised to  $T_h > T_m$  and maintained there while the left vertical face is maintained at  $T_c$ . All other faces are insulated.

Melting proceeds from right (i.e.  $X = 1$ ) towards left ( $X = 0$ ). However the process conditions are such that after sufficient time, convective heat transfer from the melt

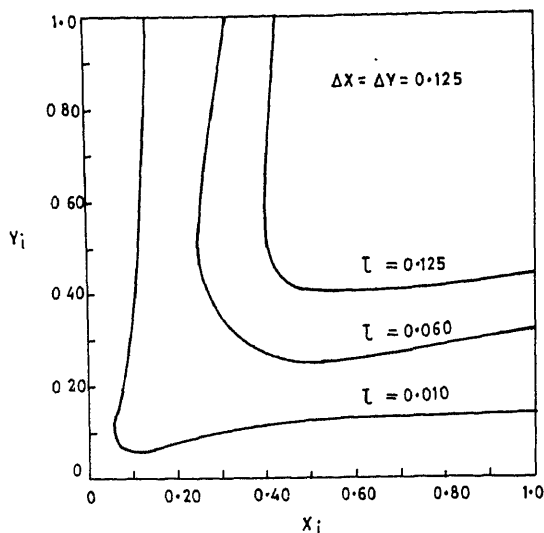


Figure 5. Evolution of interface - 2-D conduction phase change.

to the interface is balanced by the heat conduction through the solid, and a steady state is reached.

Pillay (1992) performed calculations assuming the melting situation to be two-dimensional neglecting all variations in the depthwise direction. The same conditions were assumed by Beckermann & Viskanta (1989).

Figure 6 shows the comparison of predicted and experimental interface location at steady-state for three experimental conditions:  $T^* = (T_h - T_c)/(T_h - T_m) = 0.468, 0.957, 1.935$ . Pillay's (1992) predictions were obtained with  $10 \times 10$  fixed nodes, whereas Beckermann & Viskanta (1989), who used a variable domain formulation, used  $26 \times 42$  grid nodes in liquid and solid regions each. Pillay (1992) solved the enthalpy and momentum equations simultaneously and not through a quasi-steady approximation necessitated in the variable domain method for solving the momentum equation. Considering the coarseness of the grid, Pillay's predictions are in reasonable agreement with the experiment and the predictions of Beckermann & Viskanta (1989). The departure of the interface shape from the vertical demonstrates the effect of buoyancy-induced circulation in the liquid region.

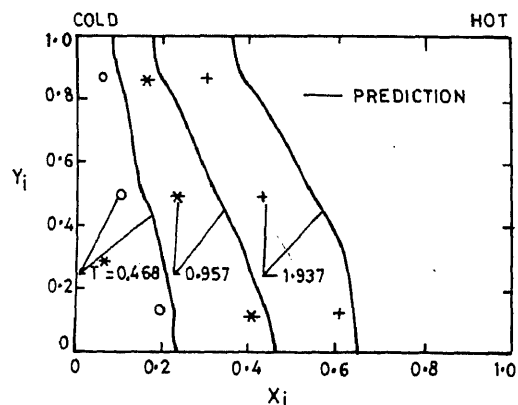


Figure 6. Interface position at steady state - 2-D phase change with convection.



## 5. Conclusions

The following are the salient features of the enthalpy formulation described in the present paper.

- (1) By generalising the temperature-enthalpy relationship, first in a continuum ( $\theta = \phi + \phi'$ ) and then adapting it to the requirements of discretization ( $\theta = \phi + \phi'' = \phi + \phi' + \theta_{pc}$ ), a calculation procedure is developed that is numerically stable and physically accurate resulting in prediction of oscillation-free, non-wavy temperature histories and interface movements. The introduction of  $\theta_{pc}$  is akin to stating that phase change takes place over a range of temperatures. However, the range here is physically estimated and variable, and is not externally imposed as done by earlier researchers (see for example, Meyer 1971 or Szekely & Themlis 1970).
- (2) The  $\phi'$  and  $\phi''$  variables are used to locate the interface locally, to suppress  $\theta_{pc}$  to zero at single phase nodes, and to enable numerical integration of the discretised equations by line-by-line algorithm without node "book-keeping". The  $\phi'$  variable is also used to calculate the liquid portion of the control volume necessary for integrating Navier-Stokes equations when convection is present.
- (3) Since the interface coordinates are calculated with respect to a local origin, rather than with respect to a fixed origin, the present method can handle interface of arbitrary shape and one that may multiply the intersect with a fixed grid line. Third-order accurate Taylor's series expansion is used to estimate the location of the interface. As such when the interface is very sharply curved, finer mesh size is required.
- (4) The present method has been extended to account for differences in liquid and solid properties, non-uniform grid spacing and convection. It is suggested that solution of equations of motion on non-staggered grid (Date 1993, 1994) enhances the convenience offered by the enthalpy formulation. The method can also be extended to complex geometries, and to the use of general curvilinear coordinates.

## Appendix A – Expressions for location of the interface

Refer to figure 3.

Type (b) intersection ( $\phi'_{k,j-1} = 0, \phi'_{k,j+1} = -1$ )

$$\Delta Y_i = (0.5 + \phi''_{kj}) \Delta Y, \quad (\text{A1})$$

$$\Delta X_i = -(0.5 + \phi''_{kj}) \left/ \left\{ \frac{\partial \phi'}{\partial X} \right|_{kj} \right\} + \left\{ (0.5 + \phi''_{kj})^2 \frac{\partial^2 \phi'}{\partial X \partial Y} \right|_{kj} \Delta Y \right\} \left/ \left\{ 2 \left( \frac{\partial \phi'}{\partial X} \right|_{kj} \right)^2 \right\}. \quad (\text{A2})$$

Type (d) intersection ( $\phi'_{k,j+1} = \phi'_{k+1,j} = -1$ )

$$\begin{aligned} \Delta Y_i = & -0.5 \Delta Y - (0.5 + \phi''_s) \left/ \left\{ \frac{\partial \phi'}{\partial Y} \right|_s \right\} \\ & + \left\{ (0.5 + \phi''_s)^2 \frac{\partial^2 \phi'}{\partial X \partial Y} \right|_s \Delta x \right\} \left/ \left\{ 2 \left( \frac{\partial \phi'}{\partial Y} \right|_s \right)^2 \right\}, \end{aligned} \quad (\text{A3})$$

$$\Delta X_i = -0.5\Delta X - (0.5 + \phi''_w) \left/ \left\{ \frac{\partial \phi'}{\partial X} \right|_w \right\} + \left\{ (0.5 + \phi''_w)^2 \frac{\partial^2 \phi'}{\partial X \partial Y} \right|_2 \Delta Y \right\} / \left\{ 2 \left( \frac{\partial \phi'}{\partial X} \right|_w \right\}^2 \quad (\text{A4})$$

Type (e) intersection ( $\phi'_{k+1,j} = \phi'_{k,j-1} = -1$ )

$\Delta Y_i = \text{RHS of (36g),}$

$\Delta X_i = \text{RHS of (A4).}$

Type (f) intersection ( $\phi'_{k,j+1} = \phi'_{k-1,j} = 0$ )

$\Delta Y_i = \text{RHS of (A3),}$

$\Delta X_i = \text{RHS of (36h).}$

### List of symbols

$C_p$	specific heat;
$H$	total enthalpy;
$H_{ps}$	pseudo enthalpy;
$K$	thermal conductivity;
$L$	characteristic length;
$S, S_x$	$\Delta\tau/\Delta X^2$ ;
$S_y$	$\Delta\tau/\Delta Y^2$ ;
$St$	Stefan number $[C_p(T_m - T_{ref})/\lambda]$ ;
$t$	time;
$T$	temperature;
$v_i$	interface velocity;
$u$	velocity;
$X$	dimensionless x coordinate ( $X/L$ );
$Y$	dimensionless y coordinate ( $Y/L$ );
$X_i, Y_i$	interface coordinates;
$\alpha$	thermal diffusivity;
$\theta$	dimensionless temperature;
$\lambda$	latent heat;
$\rho$	density;
$\tau$	dimensionless time;
$\phi$	dimensionless enthalpy;
$\phi'$	see (14);
$\phi''$	$\phi' + \theta_{pc}$ .

### Subscripts

$i$	interface;
$k, j$	node designation;
$in$	initial condition;

<i>l</i>	liquid;
<i>m</i>	melting point;
<i>pc</i>	phase-change;
<i>s</i>	solid;
<i>e, n, s, w</i>	control-volume cell-face locations.

### Superscripts

<i>n</i>	new value;
<i>o</i>	old value.

### References

- Basu B, Date A W 1987 On local vs global implicitness during the solution of melting and solidification problems. *Proc. 9th ISHMT Heat Mass Transfer Conference, Bangalore, India* paper no. HMT-57-87
- Basu B, Date A W 1988 Numerical modeling of melting and solidification problems-A review. *Sādhanā* 13: 169-213
- Beckermann C, Viskanta R 1989 Effect of sub-cooling on natural convection melting of a pure-metal. *Trans. ASME, J. Heat Transfer* 3: 416-424
- Date A W 1991 A strong enthalpy formulation for the Stefan problem. *Int. J. Heat Mass Transfer* 34: 2231-2235
- Date A W 1992 Novel strongly implicit enthalpy formulation for multidimensional Stefan problems. *Numer. Heat Transfer* B21: 231-235
- Date A W 1993 Solution of Navier-Stokes equations on non-staggered grid. *Int. J. Heat Mass Transfer* 36: 1913-1922
- Date A W 1994 A calculation procedure for prediction of heat mass and momentum transfer in elliptic flows using non-staggered grid. *Proceedings 1st ISHMT-ASME Heat Mass Transfer Conference, Bombay*
- Gadgil A, Gobin D 1984 Analysis of two-dimensional melting in rectangular enclosure in presence of convection. *Trans. ASME, J. Heat Transfer* 106: 20-26
- Huang L I, Ayyaswamy P S, Cohen I M 1991 A note on interface condition in phase-change problems. *Trans. ASME J. Heat Transfer* 113: 244-247
- Lazaridis A 1970 A numerical solution of multidimensional solidification (or melting) problems. *Int. J. Heat Mass Transfer* 13: 155-169
- Meyer G H 1971 Multidimensional Stefan problems. *SIAM J. Numer. Anal.* 8: 80-96
- Patel P D 1968 Interface condition in heat conduction problems with change of phase. *AIAA J.* 6: 2454
- Pillay V R 1992 Numerical investigation of heat transfer during process of solidification & melting with and without convection. M. Tech. Dissertation, Dept. of Mech. Eng., Indian Inst. Technol., Bombay
- Rhie C M, Chow W L 1983 A numerical study of the turbulent flow past an isolated airfoil with trailing edge separation. *AIAA J.* 21: 1525-1532
- Saitoh T 1978 Numerical method for multidimensional freezing problems in arbitrary regions. *Trans. ASME, J. Heat Transfer* 100: 294-299
- Shamsunder N 1978 Comparison of numerical methods for diffusion problems with moving boundaries In *Moving boundary problems* (eds) D G Wilson (New York: Academic Press)
- Shamsunder N, Sparrow E M 1975 Analysis of conduction phase-change via enthalpy model. *Trans. ASME, J. Heat Transfer* 97: 333-340
- Sparrow E M, Schneider G E, Pletcher R H 1988 *Handbook of numerical heat transfer* (New York: Wiley) chap. 18
- Szekely J, Thémelis N J 1970 *Rate phenomena in process metallurgy* (New York: Wiley Interscience) chap. 10
- Voller V R 1990 Fast implicit finite difference method for the analysis of phase-change problems. *Numer. Heat Transfer* B17: 155-169

- Voller V R, Cross M, Markatos N C 1987 An enthalpy method for convection diffusion phase change. *Int. J. Numer. Methods Eng.* 24: 271–284
- Voller V R, Cross M, Walton P G 1979 Assessment of weak solution numerical techniques for solving Stefan problems. In *Numerical methods in thermal problems* (eds) R W Lewis, K Morgan (London: Pineridge)
- Wilson D G, Soloman A D, Boggs P T 1978 *Moving boundary problems* (New York: Academic Press)

# A tutorial survey of reinforcement learning

S SATHIYA KEERTHI and B RAVINDRAN

Department of Computer Science and Automation,  
Indian Institute of Science, Bangalore 560 012, India

**Abstract.** This paper gives a compact, self-contained tutorial survey of reinforcement learning, a tool that is increasingly finding application in the development of intelligent dynamic systems. Research on reinforcement learning during the past decade has led to the development of a variety of useful algorithms. This paper surveys the literature and presents the algorithms in a cohesive framework.

**Keywords.** Reinforcement learning; dynamic programming; optimal control; neural networks.

## 1. Introduction

Reinforcement Learning (RL), a term borrowed from animal learning literature by Minsky (1954, 1961), refers to a class of learning tasks and algorithms in which the learning system learns an associative mapping,  $\pi : X \rightarrow A$  by maximizing a scalar evaluation (reinforcement) of its performance from the environment (user). Compared to supervised learning, in which for each  $x$  shown the environment provides the learning system with the value of  $\pi(x)$ , RL is more difficult since it has to work with much less feedback from the environment. If, at some time, given an  $x \in X$ , the learning system tries an  $a \in A$  and, the environment immediately returns a scalar reinforcement evaluation of the  $(x, a)$  pair (that indicates how far  $a$  is from  $\pi(x)$ ) then we are faced with an *immediate* RL task. A more difficult RL task is *delayed* RL, in which the environment only gives a single scalar reinforcement evaluation, collectively for  $\{(x_t, a_t)\}$ , a sequence of  $(x, a)$  pairs occurring in time during the system operation. Delayed RL tasks commonly arise in optimal control of dynamic systems and planning problems of AI. In this paper our main interest is in the solution of delayed RL problems. However, we also study immediate RL problems because methods of solving them play an useful role in the solution of delayed RL problems.

Delayed RL encompasses a diverse collection of ideas having roots in animal learning (Barto 1985; Sutton & Barto 1987), control theory (Bertsekas 1989; Kumar 1985), and AI (Dean & Wellman 1991). Delayed RL algorithms were first employed by Samuel (1959, 1967) in his celebrated work on playing checkers. However, it

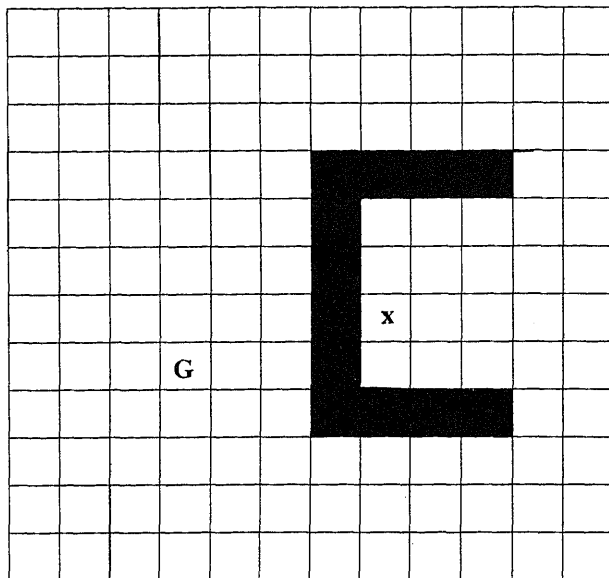


Figure 1. Navigating in a grid world.

was only much later, after the publication of Barto, Sutton and Anderson's work (Barto *et al* 1983) on a delayed RL algorithm called *adaptive heuristic critic* and its application to the control problem of pole balancing, that research on RL got off to a flying start. Watkins' *Q*-Learning algorithm (Watkins 1989) made another impact on the research. A number of significant ideas have rapidly emerged during the past five years and the field has reached a certain level of maturity. In this paper we provide a comprehensive tutorial survey of various ideas and methods of delayed RL. To avoid distractions and unnecessary clutter of notations, we present all ideas in an intuitive, not-so-rigorous fashion. In preparing this tutorial, we have obtained a lot of guidance from the works of Watkins (1989), Barto *et al* (1990, 1992), Bradtke (1994), and Barto (1992).

To illustrate the key features of a delayed RL task let us consider a simple example.

#### *Example 1 Navigating a Robot*

Figure 1 illustrates a grid world in which a robot navigates. Each blank cell on the grid is called a *state*. Shaded cells represent barriers; these are not states. Let  $X$  be the state space, i.e., the set of states. The cell marked  $G$  is the goal state. The aim is to reach  $G$  from any state in the least number of time steps. Navigation is done using four *actions*:  $A = \{N, S, E, W\}$ , the actions denoting the four possible movements along the coordinate directions.

Rules of transition are defined as follows. Suppose that the robot is in state  $x$  and action  $N$  is chosen. Then the resulting next state,  $y$  is the state directly to the north of  $x$ , if there is such a state; otherwise  $y = x$ . For instance, choosing  $W$  at the  $x$  shown in figure 1 will lead to the system staying at  $x$ . The goal state is a special case. By definition we will take it that any action taken from the goal state results in a transition back to the goal state. In more general problems, the rules of transition can be stochastic.

The robot moves at discrete (integer) time points starting from  $t = 0$ . At a time step  $t$ , when the robot is at state,  $x_t$ , we define an immediate reward<sup>1</sup> as

$$r(x_t) = \begin{cases} 0 & \text{if } x_t = G, \\ -1 & \text{otherwise.} \end{cases}$$

In effect; the robot is penalized for every time step spent at non-goal states. It is simple to verify that maximizing the *total reward* over time,

$$V(x) = \sum_{t=0}^{\infty} r(x_t)$$

is equivalent to achieving minimum time navigation from the starting state,  $x_0 = x$ . Let  $V^*(x)$  denote the maximum achievable (optimal) value of  $V(x)$ .

We are interested in finding a feedback policy,  $\pi : X \rightarrow A$  such that, if we start from any starting state and select actions using  $\pi$  then we will always reach the goal in the minimum number of time steps.

The usefulness of immediate RL methods in delayed RL can be roughly explained as follows. Typical delayed RL methods maintain  $\hat{V}$ , an approximation of the optimal function,  $V^*$ . If action  $a$  is performed at state  $x$  and state  $y$  results, then  $\hat{V}(y)$  can be taken as an (approximate) immediate evaluation of the  $(x, a)$  pair.<sup>2</sup> By solving an immediate RL problem that uses this evaluation function we can obtain a good sub-optimal policy for the delayed RL problem. We present relevant immediate RL algorithms in §2.

□

Delayed RL problems are much harder to solve than immediate RL problems for the following reason. Suppose, in example 1, performance of a sequence of actions, selected according to some policy, leads the robot to the goal. To improve the policy using the experience, we need to evaluate the goodness of each action performed. But the total reward obtained gives only the cumulative effect of all actions performed. Some scheme must be found to reasonably apportion the cumulative evaluation to the individual actions. This is referred to as the *temporal credit assignment problem*. (In the previous paragraph we have already given a hint of how delayed RL methods do temporal credit assignment.)

Dynamic programming (DP) (Bertsekas 1989; Ross 1983) is a well-known tool for solving problems such as the one in example 1. It is an off-line method that requires the availability of a complete model of the environment. But the concerns of delayed RL are very different. To see this clearly let us return to example 1 and impose the requirement that the robot has no knowledge of the environment and that the only way of learning is by on-line experience of trying various actions<sup>3</sup> and thereby visiting many states. Delayed RL algorithms are particularly meant for such situations and have the following general format.

### Delayed RL algorithm

*Initialize the learning system.*

*Repeat*

<sup>1</sup>Sometimes  $r$  is referred to as the primary reinforcement. In more general situations,  $r$  is a function of  $x_t$  as well as  $a_t$ , the action at time step  $t$ .

<sup>2</sup>An optimal action at  $x$  is one that gives the maximum value of  $V^*(y)$ .

<sup>3</sup>During learning this is usually achieved by using a (stochastic) exploration policy for choosing actions. Typically the exploration policy is chosen to be totally random at the beginning of learning and made to approach an optimal policy as learning nears completion.

1. With the system at state  $x$ , choose an action  $a$  according to an exploration policy and apply it to the system.
2. The environment returns a reward,  $r$ , and also yields the next state,  $y$ .
3. Use the experience,  $(x, a, r, y)$  to update the learning system.
4. Set  $x := y$ .

Even when a model of the environment is available, it is often advantageous to avoid an off-line method such as DP and instead use a delayed RL algorithm. This is because, in many problems the state space is very large; while a DP algorithm operates with the entire state space, a delayed RL algorithm only operates on parts of the state space that are most relevant to the system operation. When a model is available, delayed RL algorithms can employ simulation mode of operation instead of on-line operation so as to speed-up learning and avoid doing experiments using hardware. In this paper, we will use the term, *real time operation* to mean that either on-line operation or simulation mode of operation is used.

In most applications, representing functions such as  $V^*$  and  $\pi$  exactly is infeasible. A better alternative is to employ parametric function approximators, e.g., connectionist networks. Such approximators must be suitably chosen for use in a delayed RL algorithm. To clarify this, let us take  $V^*$  for instance and consider a function approximator,  $\hat{V}(\cdot; w) : X \rightarrow R$ , for it. Here  $R$  denotes the real line and  $w$  denotes the vector of parameters of the approximator that is to be learnt so that  $\hat{V}$  approximates  $V^*$  well. Usually, at step 3 of the delayed RL algorithm, the learning system uses the experience to come up with a direction,  $\eta$  in which  $\hat{V}(x; w)$  has to be changed for improving performance. Given a step size,  $\beta$ , the function approximator must alter  $w$  to a new value,  $w^{\text{new}}$  so that

$$\hat{V}(x; w^{\text{new}}) = \hat{V}(x; w) + \beta\eta \quad (1)$$

For example, in multilayer perceptrons (Hertz *et al* 1991)  $w$  denotes the set of weights and thresholds in the network and, their updating can be carried out using back-propagation so as to achieve (1). In the rest of the paper we will denote the updating process in (1) as

$$\hat{V}(x; w) := \hat{V}(x; w) + \beta\eta \quad (2)$$

and refer to it as a *learning rule*.

The paper is organized as follows. Section 2 discusses immediate RL. In §3 we formulate Delayed RL problems and mention some basic results. Methods of estimating total reward are discussed in §4. These methods play an important role in delayed RL algorithms. DP techniques and delayed RL algorithms are presented in §5. Section 6 addresses various practical issues. We make a few concluding remarks in §7.

## 2. Immediate reinforcement learning

Immediate RL refers to the learning of an associative mapping,  $\pi : X \rightarrow A$  given a reinforcement evaluator. To learn, the learning system interacts in a closed loop with the environment. At each time step, the environment chooses an  $x \in X$  and, the learning system uses its function approximator,  $\hat{\pi}(\cdot; w)$  to select an action:



$a = \hat{\pi}(x; w)$ . Based on both  $x$  and  $a$ , the environment returns an evaluation or "reinforcement",  $r(x, a) \in R$ . Ideally, the learning system has to adjust  $w$  so as to produce the maximum possible  $r$  value for each  $x$ ; in other words, we would like  $\hat{\pi}$  to solve the parametric global optimization problem,

$$r(x, \hat{\pi}(x; w)) = r^*(x) \stackrel{\text{def}}{=} \max_{a \in A} r(x, a) \quad \forall x \in X \quad (3)$$

Supervised learning is a popular paradigm for learning associative mappings (Hertz *et al* 1991). In supervised learning, for each  $x$  shown the supervisor provides the learning system with the value of  $\pi(x)$ . Immediate RL and supervised learning differ in the following two important ways.

- In supervised learning, when an  $x$  is shown and the supervisor provides  $a = \pi(x)$ , the learning system forms the directed information,  $\eta = a - \hat{\pi}(x; w)$  and uses the learning rule:  $\hat{\pi}(x; w) := \hat{\pi}(x; w) + \alpha\eta$ , where  $\alpha$  is a (positive) step size. For immediate RL such directed information is not available and so it has to employ some strategy to obtain such information.
- In supervised learning, the learning system can simply check if  $\eta = 0$  and hence decide whether the correct map value has been formed by  $\hat{\pi}$  at  $x$ . However, in immediate RL, such a conclusion on correctness cannot be made without exploring the values of  $r(x, a)$  for all  $a$ .

Therefore, immediate RL problems are much more difficult to solve than supervised learning problems.

A number of immediate RL algorithms have been described in the literature. Stochastic learning automata algorithms (Narendra & Thathachar 1989) deal with the special case in which  $X$  is a singleton,  $A$  is a finite set, and  $r \in [0, 1]$ .<sup>4</sup> The Associative Reward-Penalty ( $AR-P$ ) algorithm (Barto & Anandan 1985; Barto *et al* 1985; Barto & Jordan 1987; Mazzoniet *al* 1990) extends the learning automata ideas to the case where  $X$  is a finite set. Williams (1986, 1987) has proposed a class of immediate RL methods and has presented interesting theoretical results. Gullapalli (1990, 1992a) has developed algorithms for the general case in which  $X$ ,  $A$  are finite-dimensional real spaces and  $r$  is real valued. Here we will discuss only algorithms which are most relevant to, and useful in delayed RL.

One simple way of solving (3) is to take one  $x$  at a time, use a global optimization algorithm (e.g., complete enumeration) to explore the  $A$  space and obtain the correct  $a$  for the given  $x$ , and then make the function approximator learn this  $(x, a)$  pair. However, such an idea is not used for the following reason. In most situations where immediate RL is used as a tool (e.g., to approximate a policy in delayed RL), the learning system has little control over the choice of  $x$ . When, at a given  $x$ , the learning system chooses a particular  $a$  and sends it to the environment for evaluation, the environment not only sends a reinforcement evaluation but also alters the  $x$  value. Immediate RL seeks approaches which are appropriate to these situations.

Let us first consider the case in which  $A$  is a finite set:  $A = \{a^1, a^2, \dots, a^m\}$ . Let  $R^m$  denote the  $m$ -dimensional real space. The function approximator,  $\hat{\pi}$  is usually

<sup>4</sup>Stochastic Learning Automata algorithms can also be used when  $X$  is not a singleton, by employing teams of co-operating automata. For more details on such algorithms see Narendra & Thathachar (1989).

formed as a composition of two functions: a function approximator,  $g(\cdot; w) : X \rightarrow R^m$  and a fixed function,  $M : R^m \rightarrow A$ . The idea behind this set-up is as follows. For each given  $x$ ,  $z = g(x; w) \in R^m$  gives a vector of merits of the various  $a^i$  values. Let  $z_k$  denote the  $k$ -th component of  $z$ . Given the merit vector  $z$ ,  $a = M(z)$  is formed by the max selector,

$$a = a^k \quad \text{where} \quad z_k = \max_{1 \leq i \leq m} z_i \quad (4)$$

Let us now come to the issue of learning (i.e., choosing a  $w$ ). At some stage, let  $x$  be the input,  $z$  be the merit vector returned by  $g$ , and  $a^k$  be the action having the largest merit value. The environment returns the reinforcement,  $r(x, a^k)$ . In order to learn we need to evaluate the goodness of  $z^k$  (and therefore, the goodness of  $a^k$ ). Obviously, we cannot do this using existing information. We need an estimator, call it  $\hat{r}(x; v)$ , that provides an estimate of  $r^*(x)$ . The difference,  $r(x, a^k) - \hat{r}(x; v)$  is a measure of the goodness of  $a^k$ . Then a simple learning rule is

$$g_k(x; w) := g_k(x; w) + \alpha(r(x, a^k) - \hat{r}(x; v)) \quad (5)$$

where  $\alpha$  is a small (positive) step size.

Learning  $\hat{r}$  requires that all members of  $A$  are evaluated by the environment at each  $x$ . Clearly, the max selector, (4) is not suitable for such exploration. For instance, if at some stage of learning, for some  $x$ ,  $g$  assigns the largest merit to a wrong action, say  $a^k$ , and  $\hat{r}$  gives, by mistake, a value smaller than  $r(x, a^k)$ , then no action other than  $a^k$  is going to be generated by the learning system at the given  $x$ . So we replace (4) by a controlled stochastic action selector that generates actions randomly when learning begins and approaches (4) as learning is completed. A popular stochastic action selector is based on the Boltzmann distribution,

$$p_i(x) \stackrel{\text{def}}{=} \text{Prob}\{a = a^i | x\} = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (6)$$

where  $T$  is a nonnegative real parameter (temperature) that controls the stochasticity of the action selector. For a given  $x$  the expected reinforcement of the action selector is

$$\tilde{r}(x) \stackrel{\text{def}}{=} E(r(x, a) | x) = \sum_i p_i(x) r(x, a^i)$$

As  $T \rightarrow 0$  the stochastic action selector approaches the max selector, (4), and,  $\tilde{r}(x) \rightarrow r^*(x)$ . Therefore we train  $\hat{r}$  to approximate  $\tilde{r}$  (instead of  $r^*$ ). This is easy to do because, for any fixed value of  $T$ ,  $\tilde{r}$  can be estimated by the average of the performance of the stochastic action selector over time. A simple learning rule that achieves this is

$$\hat{r}(x; v) := \hat{r}(x; v) + \beta(r(x, a) - \hat{r}(x; v)) \quad (7)$$

where  $\beta$  is a small (positive) step size.

**Remark** Two important comments should be made regarding the convergence of learning rules such as (7) (we will come across many such learning rules later) which are designed to estimate an expectation by averaging over time.

- Even if  $\hat{r} \equiv \tilde{r}$ ,  $r(x, a) - \hat{r}(x; v)$  can be non-zero and even large in size. This is because  $a$  is only an instance generated by the distribution,  $p(x)$ . Therefore, to avoid unlearning as  $\hat{r}$  comes close to  $\tilde{r}$ , the step size,  $\beta$  must be controlled properly. The value of  $\beta$  may be chosen to be 1 when learning begins, and then slowly decreased to 0 as learning progresses.

- For good learning to take place, the sequence of  $x$  values at which (7) is carried out must be such that it covers all parts of the space,  $X$  as often as possible. Of course, when the learning system has no control over the choice of  $x$ , it can do nothing to achieve such an exploration. To explore, the following is usually done. Learning is done over a number of *trials*. A trial consists of beginning with a random choice of  $x$  and operating the system for several time steps. At any one time step, the system is at some  $x$  and the learning system chooses an action,  $a$  and learns using (7). Depending on  $x$ ,  $a$  and the rules of the environment a new  $x$  results and the next time step begins. Usually, when learning is repeated over multiple trials, the  $X$  space is thoroughly explored.

Let us now consider the case in which  $A$  is continuous, say a finite dimensional real space. The idea of using merit values is not suitable. It is better to directly deal with a function approximator,  $h(\cdot; w)$  from  $X$  to  $A$ . In order to do exploration a controlled random perturbation,  $\eta$  is added to  $h(x; w)$  to form  $a = \hat{\pi}(x)$ . A simple choice is to take  $\eta$  to be a Gaussian with zero mean and having a standard deviation,  $\sigma(T)$  that satisfies:  $\sigma(T) \rightarrow 0$  as  $T \rightarrow \infty$ . The setting-up and training of the reinforcement estimator,  $\hat{r}$  is as in the case when  $A$  is discrete. The function approximator,  $h$  can adopt the following learning rule:

$$h(x; w) := h(x; w) + \alpha(r(x, a) - \hat{r}(x; w))\eta \quad (8)$$

where  $\alpha$  is a small (positive) step size. In problems where a bound on  $r^*$  is available, this bound can be suitably employed to guide exploration, i.e., choose  $\sigma$  (Gullapalli 1990).

Jordan & Rumelhart (1990) have suggested a method of 'forward models' for continuous action spaces. If  $r$  is a known differentiable function, then a simple, deterministic learning law based on gradient ascent can be given to update  $\hat{\pi}$ :

$$\hat{\pi}(x; w) := \hat{\pi}(x; w) + \alpha \frac{\partial r(x, a)}{\partial a} \quad (9)$$

If  $r$  is not known, Jordan and Rumelhart suggest that it is learnt using on-line data, and (9) be used using this learnt  $r$ . If for a given  $x$ , the function  $r(x, \cdot)$  has local maxima then the  $\hat{\pi}(x)$  obtained using learning rule, (9) may not converge to  $\pi(x)$ . Typically this is not a serious problem. The stochastic approach discussed earlier does not suffer from local maxima problems. However, we should add that, because the deterministic method explores in systematic directions and the stochastic method explores in random directions, the former is expected to be much faster. The comparison is very similar to the comparison of deterministic and stochastic techniques of continuous optimization.

### 3. Delayed reinforcement learning

Delayed RL concerns the solution of stochastic optimal control problems. In this section we discuss the basics of such problems. Solution methods for delayed RL will be presented in §4 and §5. In these three sections we will mainly consider problems in which the state and control spaces are finite sets. This is because the main issues and solution methods of delayed RL can be easily explained for such problems. We will deal with continuous state and/or action spaces briefly in §5.

Consider a discrete-time stochastic dynamic system with a finite set of states,  $X$ . Let the system begin its operation at  $t = 0$ . At time  $t$  the *agent (controller)* observes state<sup>5</sup>  $x_t$  and, selects (and performs) action  $a_t$  from a finite set,  $A(x_t)$ , of possible actions. Assume that the system is Markovian and stationary, i.e.,

$$\begin{aligned} \text{Prob}\{x_{t+1} = y \mid x_0, a_0, x_1, a_1, \dots, x_t = x, a_t = a\} \\ = \text{Prob}\{x_{t+1} = y \mid x_t = x, a_t = a\} \stackrel{\text{def}}{=} P_{xy}(a) \end{aligned}$$

A *policy* is a method adopted by the agent to choose actions. The objective of the decision task is to find a policy that is optimal according to a well defined sense, described below. In general, the action specified by the agent's policy at some time can depend on the entire past history of the system. Here we restrict attention to policies that specify actions based only on the current state of the system. A deterministic policy,  $\pi$  defines, for each  $x \in X$  an action  $\pi(x) \in A(x)$ . A stochastic policy,  $\pi$  defines, for each  $x \in X$  a probability distribution on the set of feasible actions at  $x$ , i.e., it gives the values of  $\text{Prob}\{\pi(x) = a\}$  for all  $a \in A(x)$ . For the sake of keeping the notations simple we consider only deterministic policies in this section. All ideas can be easily extended to stochastic policies using appropriate detailed notations.

Let us now precisely define the optimality criterion. While at state  $x$ , if the agent performs action  $a$ , it receives an immediate *payoff* or *reward*,  $r(x, a)$ . Given a policy  $\pi$  we define the *value function*,  $V^\pi : X \rightarrow R$  as follows:

$$V^\pi(x) = E\left\{\sum_{t=0}^{\infty} \gamma^t r(x_t, \pi(x_t)) \mid x_0 = x\right\} \quad (10)$$

Here future rewards are discounted by a factor  $\gamma \in [0, 1)$ . The case  $\gamma = 1$  is avoided only because it leads to some difficulties associated with the existence of the summation in (10). Of course, these difficulties can be handled by putting appropriate assumptions on the problem solved. But, to avoid unnecessary distraction we do not go into the details; see (Bradtke 1994; Bertsekas & Tsitsiklis 1989).

The expectation in (10) should be understood as

$$V^\pi(x) = \lim_{N \rightarrow \infty} E\left\{\sum_{t=0}^{N-1} \gamma^t r(x_t, \pi(x_t)) \mid x_0 = x\right\}$$

where the probability with which a particular state sequence,  $\{x_t\}_{t=0}^{N-1}$  occurs is taken in an obvious way using  $x_0 = x$  and repeatedly employing  $\pi$  and  $P$ . We wish to maximize the value function:

$$V^*(x) = \max_{\pi} V^\pi(x) \quad \forall x \quad (11)$$

$V^*$  is referred to as the optimal value function. Because  $0 \leq \gamma < 1$ ,  $V^\pi(x)$  is bounded. Also, since the number of  $\pi$ 's is finite  $V^*(x)$  exists.

<sup>5</sup>If the state is not completely observable then a method that uses the observable states and retains past information has to be used; see (Bacharach 1991; Bacharach 1992; Chrisman 1992; Mozer & Bacharach 1990a, 1990b; Whitehead & Ballard 1990).

How do we define an optimal policy,  $\pi^*$ ? For a given  $x$  let  $\pi^{x,*}$  denote a policy that achieves the maximum in (11). Thus we have a collection of policies,  $\{\pi^{x,*} : x \in X\}$ . Now  $\pi^*$  is defined by picking only the first action from each of these policies:

$$\pi^*(x) = \pi^{x,*}(x), \quad x \in X$$

It turns out that  $\pi^*$  achieves the maximum in (11) for every  $x \in X$ . In other words,

$$V^*(x) = V^{\pi^*}(x), \quad x \in X \quad (12)$$

This result is easy to see if one looks at Bellman's optimality equation – an important equation that  $V^*$  satisfies:

$$V^*(x) = \max_{a \in A(x)} \left[ r(x, a) + \gamma \sum_{y \in X} P_{xy}(a) V^*(y) \right] \quad (13)$$

The fact that  $V^*$  satisfies (13) can be explained as follows. The term within square brackets on the right hand side is the total reward that one would get if action  $a$  is chosen at the first time step and then the system performs optimally in all future time steps. Clearly, this term cannot exceed  $V^*(x)$  since that would violate the definition of  $V^*(x)$  in (11); also, if  $a = \pi^{x,*}(x)$  then this term should equal  $V^*(x)$ . Thus (13) holds. It also turns out that  $V^*$  is the unique function from  $X$  to  $R$  that satisfies (13) for all  $x \in X$ . This fact, however, requires a non-trivial proof; details can be found in (Ross 1983; Bertsekas 1989; Bertsekas & Tsitsiklis 1989).

The above discussion also yields a mechanism for computing  $\pi^*$  if  $V^*$  is known:

$$\pi^*(x) = \arg \max_{a \in A(x)} \left[ r(x, a) + \gamma \sum_{y \in X} P_{xy}(a) V^*(y) \right]$$

A difficulty with this computation is that the system model, i.e., the function,  $P_{xy}(a)$  must be known. This difficulty can be overcome if, instead of the  $V$ -function we employ another function called the  $Q$ -function. Let  $\mathcal{U} = \{(x, a) : x \in X, a \in A(x)\}$ , the set of feasible (state, action) pairs. For a given policy  $\pi$ , let us define  $Q^\pi : \mathcal{U} \rightarrow R$  by

$$Q^\pi(x, a) = r(x, a) + \gamma \sum_{y \in X} P_{xy}(a) V^\pi(y) \quad (14)$$

Thus  $Q^\pi(x, a)$  denotes the total reward obtained by choosing  $a$  as the first action and then following  $\pi$  for all future time steps. Let  $Q^* = Q^{\pi^*}$ . By Bellman's optimality equation and (12) we get

$$V^*(x) = \max_{a \in A(x)} [Q^*(x, a)] \quad (15)$$

It is also useful to rewrite Bellman's optimality equation using  $Q^*$  alone:

$$Q^*(x, a) = r(x, a) + \gamma \sum_{y \in X} P_{xy}(a) \left\{ \max_{b \in A(y)} Q^*(y, b) \right\} \quad (16)$$

Using  $Q^*$  we can compute  $\pi^*$ :

$$\pi^*(x) = \arg \max_{a \in A(x)} [Q^*(x, a)] \quad (17)$$

Thus, if  $Q^*$  is known then  $\pi^*$  can be computed without using a system model. This advantage of the  $Q$ -function over the  $V$ -function will play a crucial role in §5 for deriving a model-free delayed RL algorithm called  $Q$ -Learning (Watkins 1989).

Let us now consider a few examples that give useful hints for problem formulation. These examples are also commonly mentioned in the RL literature.

*Example 2 Navigating a robot with dynamics*

In example 1 the robot is moved from one cell to another like the way pieces are moved in a chess board. True robot motions, however, involve dynamics; the effects of velocity and acceleration need to be considered. In this example we will include dynamics in a crude way, one that is appropriate to the grid world. Let  $h_t$  and  $v_t$  denote the horizontal and vertical coordinates of the cell occupied by the robot at time  $t$ , and,  $\dot{h}_t$  and  $\dot{v}_t$  denote the velocities. The vector,  $(h_t, v_t, \dot{h}_t, \dot{v}_t)$  denotes the system state at time  $t$ ; each one of the four components is an integer. The goal state is  $x^G = (h^G, v^G, 0, 0)$  where  $(h^G, v^G)$  is the coordinate vector of the goal cell  $G$ . In other words, the robot has to come to rest at  $G$ . Let  $\dot{h}_{\max}$  and  $\dot{v}_{\max}$  be limits on velocity magnitudes. Thus the state space is given by

$$\tilde{X} = \{x = (h, v, \dot{h}, \dot{v}) \mid \begin{array}{l} (h, v) \text{ is a blank cell,} \\ |\dot{h}| \leq \dot{h}_{\max}, \text{ and } |\dot{v}| \leq \dot{v}_{\max} \end{array}\}$$

We will also include an extra state,  $f$  called failure state to denote situations where a barrier (shaded) cell is entered or a velocity limit is exceeded. Thus

$$X = \tilde{X} \cup \{f\}$$

The accelerations <sup>6</sup> along the horizontal and vertical directions, respectively  $a^h$  and  $a^v$ , are the actions. To keep  $h$  and  $v$  as integers let us assume that each of the accelerations takes only even integer values. Let  $a_{\max}$  be a positive even integer that denotes the limit on the magnitude of accelerations. Thus  $a = (a^h, a^v)$  is an admissible action if each of  $a^h$  and  $a^v$  is an even integer lying in  $[-a_{\max}, a_{\max}]$ .

As in example 1 state transitions are deterministic. They are defined as follows. If barrier cells and velocity limits are not present, then application of action  $(a^h, a^v)$  at  $x_t = (h_t, v_t, \dot{h}_t, \dot{v}_t)$  will lead to the next state  $x'_{t+1} = (h'_{t+1}, v'_{t+1}, \dot{h}'_{t+1}, \dot{v}'_{t+1})$  given by

$$\begin{aligned} h'_{t+1} &= h_t + \dot{h}_t + a^h/2, & v'_{t+1} &= v_t + \dot{v}_t + a^v/2 \\ \dot{h}'_{t+1} &= \dot{h}_t + a^h, & \dot{v}'_{t+1} &= \dot{v}_t + a^v \end{aligned}$$

Let  $C$  denote the curve in the grids world resulting during the transition from  $(h_t, v_t)$  at time  $t$  to  $(h'_{t+1}, v'_{t+1})$  at time  $(t+1)$ , i.e., the solution of the differential equations:  $d^2h/d\tau^2 = a^h$ ,  $d^2v/d\tau^2 = a^v$ ,  $\tau \in [t, t+1]$ ,  $h(t) = h_t$ ,  $dh/d\tau|_{\tau=t} = \dot{h}_t$ ,  $v(t) = v_t$ ,  $dv/d\tau|_{\tau=t} = \dot{v}_t$ . If, either  $C$  cuts across a barrier cell or  $(\dot{h}'_{t+1}, \dot{v}'_{t+1})$  is an inadmissible velocity vector, then we say failure has occurred during transition. Thus state transitions are defined as

$$x_{t+1} = \begin{cases} f & \text{if } x_t = f \\ f & \text{if failure occurs during transition} \\ x^G & \text{if } x^t = x^G \\ x'_{t+1} & \text{otherwise} \end{cases}$$

<sup>6</sup>Negative acceleration will mean deceleration.

The primary aim is to avoid failure. Next, among all failure-avoiding trajectories we would like to choose the trajectory which reaches the goal state,  $x^G = (h^G, v^G, 0, 0)$  in as few time steps as possible. These aims are met if we define

$$r(x, a) = \begin{cases} -1 & \text{if } x = f, \\ 1 & \text{if } x = x^G, \\ 0 & \text{otherwise.} \end{cases}$$

The following can be easily checked.

- $V^*(x) < 0$  iff there does not exist a trajectory starting from  $x$  that avoids failure.
- $V^*(x) = 0$  iff, starting from  $x$ , there exists a failure-avoiding trajectory, but there does not exist a trajectory that reaches  $G$ .
- $V^*(x) > 0$  iff, starting from  $x$ , there exists a failure-avoiding trajectory that also reaches  $G$ ; also, an optimal policy  $\pi^*$  leads to the generation of a trajectory that reaches  $G$  in the fewest number of steps from  $x$  while avoiding failure.

□

### Example 3 Playing backgammon

Consider a game of backgammon (Magriel 1976) between players A and B. Let us look at the game from A's perspective, assuming that B follows a fixed policy. Now A can make a decision on a move only when the current board pattern as well as its dice roll are known. Therefore a state consists of a (board pattern, dice roll) pair. Each action consists of a set of marker movements. State transition is defined as follows.

- A moves its markers in accordance with the chosen action. This step is deterministic, and results in a new board pattern.
- B rolls the dice. This step is stochastic.
- B moves its markers according to its policy. This step can be deterministic or stochastic depending on the type of B's policy.
- A rolls the dice. This step is stochastic.

The set of states that correspond to A's win is the set of goal states,  $G$  to be reached. We can define the reward as:  $r(x, a) = 1$  if  $x$  is a goal state; and  $r(x, a) = 0$  otherwise. If  $\gamma = 1$ , then for a given policy, say  $\pi$ , the value function  $V^\pi(x)$  will denote the probability that A will win from that state.

□

**Example 4 Pole balancing** We now deviate from our problem formulation and present an example that involves continuous state/action spaces. A standard problem for learning controllers is that of balancing an inverted pendulum pivoted on a trolley, a problem similar to that of balancing a stick on one's hand (Barto *et al* 1983). The system comprises a straight horizontal track, like a railway track, with a carriage free to move along it. On the carriage is an axis, perpendicular to the track and pointing out to the side, about which a pendulum is free to turn. The controller's task is to keep the pendulum upright, by alternately pulling and pushing

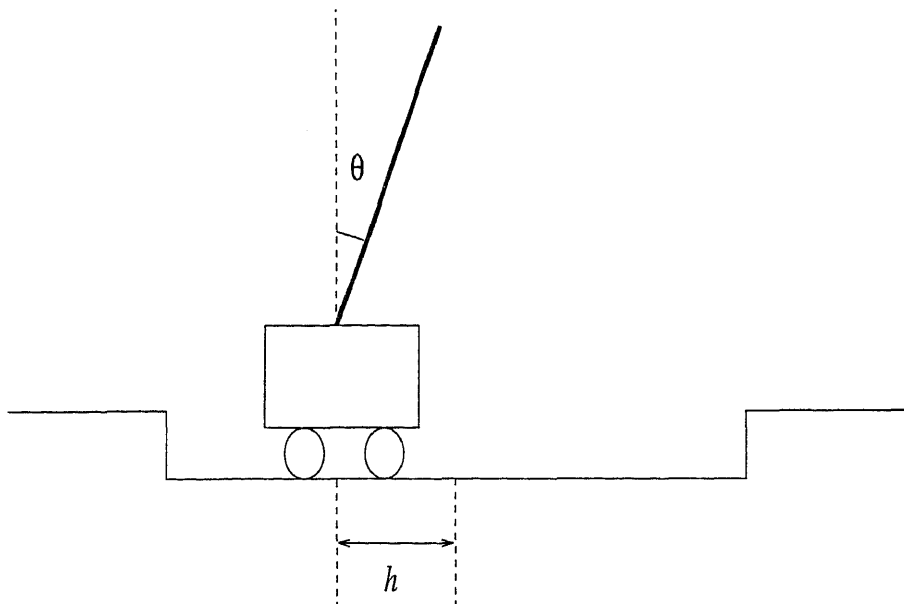


Figure 2. Pole balancing.

the carriage along the track. Let  $h$  and  $\theta$  be as shown in figure 2. We say balancing has failed if anyone of the following inequalities is violated:

$$h \leq h_{\max}, \quad h \geq -h_{\max}, \quad \theta \leq \theta_{\max}, \quad \theta \geq -\theta_{\max}$$

where  $h_{\max}$  and  $\theta_{\max}$  are specified bounds on the magnitudes of  $h$  and  $\theta$ . The aim is to balance without failure for as long a time as possible.

The state of the system is the 4-tuple,  $(h, \dot{h}, \theta, \dot{\theta})$ , where  $\dot{h}$  and  $\dot{\theta}$  are the time derivatives of  $h$  and  $\theta$  respectively. The action is the force applied to the carriage. It takes real values in the interval,  $[-F_{\max}, F_{\max}]$ . To simplify the problem solution, sometimes the action space is taken to be  $\{-F_{\max}, F_{\max}\}$  (Michie & Chambers 1968; Barto *et al* 1983; Anderson 1989). A discrete time formulation of the problem is obtained by cutting continuous time (non-negative real line) into uniform time intervals, each of duration  $\Delta$ , and taking the applied force to be constant within each interval.<sup>7</sup> The state of the system at the continuous time instant,  $t\Delta$  is taken to be  $x_t$ , the discrete time state at the  $t$ -th time step. The mechanical dynamics of the system defines state transition, except for one change: once failure occurs, we will assume, for the sake of consistent problem formulation, that the system stays at failure for ever.

As in example 2 we will take the state space to be  $X = \tilde{X} \cup \{f\}$ , where

$$\tilde{X} = \{x = (h, \dot{h}, \theta, \dot{\theta}) \mid -h_{\max} \leq h \leq h_{\max}, \quad -\theta_{\max} \leq \theta \leq \theta_{\max}\}$$

and  $f$  is the failure state that collectively represents all states not in  $\tilde{X}$ . Since the

<sup>7</sup>This constant is the action for the time step corresponding to that interval.



aim is to avoid failure, we choose

$$r(x, a) = \begin{cases} -1 & \text{if } x = f, \\ 0 & \text{otherwise.} \end{cases}$$

□

#### 4. Methods of estimating $V^\pi$ and $Q^\pi$

Delayed RL methods use a knowledge of  $V^\pi$  ( $Q^\pi$ ) in two crucial ways: (1) the optimality of  $\pi$  can be checked by seeing if  $V^\pi$  ( $Q^\pi$ ) satisfies Bellman's optimality equation; and (2) if  $\pi$  is not optimal then  $V^\pi$  ( $Q^\pi$ ) can be used to improve  $\pi$ . We will elaborate on these details in the next section. In this section we discuss, in some detail, methods of estimating  $V^\pi$  for a given policy,  $\pi$ . (Methods of estimating  $Q^\pi$  are similar and so we will deal with them briefly at the end of the section.) Our aim is to find  $\hat{V}(\cdot; v)$ , a function approximator that estimates  $V^\pi$ . Much of the material in this section is taken from the works of Watkins (1989), Sutton (1984, 1988) and Jaakkola *et al* (1994).

To avoid clumsiness we employ some simplifying notations. Since  $\pi$  is fixed we will omit the superscript from  $V^\pi$  and so call it  $V$ . We will refer to  $r(x_t, \pi(x_t))$  simply as  $r_t$ . If  $p$  is a random variable, we will use  $p$  to denote both, the random variable as well as an instance of the random variable.

A simple approximation of  $V(x)$  is the  $n$ -step truncated return,

$$V^{[n]}(x) = \sum_{\tau=0}^{n-1} \gamma^\tau r_\tau, \quad \hat{V}(x; v) = E(V^{[n]}(x)) \quad (18)$$

(Here it is understood that  $x_0 = x$ . Thus, throughout this section  $\tau$  will denote the number of time steps elapsed after the system passed through state  $x$ . It is for stressing this point that we have used  $\tau$  instead of  $t$ . In a given situation, the use of time – is it 'actual system time' or 'time relative to the occurrence of  $x$ ' – will be obvious from the context.) If  $r_{\max}$  is a bound on the size of  $r$  then it is easy to verify that

$$\max_x |\hat{V}(x; v) - V(x)| \leq \frac{\gamma^n r_{\max}}{(1 - \gamma)} \quad (19)$$

Thus, as  $n \rightarrow \infty$ ,  $\hat{V}(x; v)$  converges to  $V(x)$  uniformly in  $x$ .

But (18) suffers from an important drawback. The computation of the expectation requires the complete enumeration of the probability tree of all possible states reachable in  $n$  time steps. Since the breadth of this tree may grow very large with  $n$ , the computations can become very burdensome. One way of avoiding this problem is to set

$$\hat{V}(x; v) = V^{[n]}(x) \quad (20)$$

where  $V^{[n]}(x)$  is obtained via either Monte-Carlo simulation or experiments on the real system (the latter choice is the only way to systems for which a model is unavailable.) The approximation, (20) suffers from a different drawback. Because the breadth of the probability tree grows with  $n$ , the variance of  $V^{[n]}(x)$  also grows with  $n$ . Thus  $\hat{V}(x; v)$  in (20) will not be a good approximation of  $E(V^{[n]}(x))$  unless

it is obtained as an average over a large number of trials.<sup>8</sup> Averaging is achieved if we use a learning rule (similar to (7)):

$$\hat{V}(x; v) := \hat{V}(x; v) + \beta [V^{[n]}(x) - \hat{V}(x; v)] \quad (21)$$

where  $\beta \in (0, 1)$  is a small step size. Learning can begin with a random choice of  $v$ . Eventually, after a number of trials, we expect the  $\hat{V}$  resulting from (21) to satisfy (19).

In the above approach, an approximation of  $V$ ,  $\hat{V}$  is always available. Therefore, an estimate that is *more appropriate than*  $V^{[n]}(x)$  is the *corrected  $n$ -step truncated return*,

$$V^{(n)}(x) = \sum_{\tau=0}^{n-1} \gamma^\tau r_\tau + \gamma^n \hat{V}(x_n; v) \quad (22)$$

where  $x_n$  is the state that occurs  $n$  time steps after the system passed through state  $x$ . Let us do some analysis to justify this statement.

First, consider the ideal learning rule,

$$\hat{V}(x; v) := E(V^{(n)}(x)) \quad \forall x \quad (23)$$

Suppose  $v$  gets modified to  $v_{\text{new}}$  in the process of satisfying (23). Then, similar to (19) we can easily derive

$$\max_x |\hat{V}(x; v_{\text{new}}) - V(x)| \leq \gamma^n \max_x |\hat{V}(x; v) - V(x)|$$

Thus, as we go through a number of learning steps we achieve  $\hat{V} \rightarrow V$ . Note that this convergence is achieved even if  $n$  is fixed at a small value, say  $n = 1$ . On the other hand, for a fixed  $n$ , the learning rule based on  $V^{[n]}$ , i.e., (18), is only guaranteed to achieve the bound in (19). *Therefore, when a system model is available it is best to choose a small  $n$ , say  $n = 1$ , and employ (23).*

Now suppose that, either a model is unavailable or (23) is to be avoided because it is expensive. In this case, a suitable learning rule that employs  $V^{(n)}$  and uses real-time data is:

$$\hat{V}(x; v) := \hat{V}(x; v) + \beta [V^{(n)}(x) - \hat{V}(x; v)] \quad (24)$$

Which is better: (21) or (24)? There are two reasons as to why (24) is better.

- Suppose  $\hat{V}$  is a good estimate of  $V$ . Then a small  $n$  makes  $V^{(n)}$  ideal:  $V^{(n)}(x)$  has a mean close to  $V(x)$  and it also has a small variance. Small variance means that (24) will lead to fast averaging and hence fast convergence of  $\hat{V}$  to  $V$ . On the other hand  $n$  has to be chosen large for  $V^{[n]}(x)$  to have a mean close to  $V(x)$ ; but then,  $V^{[n]}(x)$  will have a large variance and (21) will lead to slow averaging.
- If  $\hat{V}$  is not a good estimate of  $V$  then both  $V^{(n)}$  and  $V^{[n]}$  will require a large  $n$  for their means to be good. If a large  $n$  is used, the difference between  $V^{(n)}$  and  $V^{[n]}$ , i.e.,  $\gamma^n \hat{V}$  is negligible and so both (21) and (24) will yield similar performance.

<sup>8</sup>As already mentioned, a trial consists of starting the system at a random state and then running the system for a number of time steps.

The above discussion implies that it is better to employ  $V^{(n)}$  than  $V^{[n]}$ . It is also clear that, when  $V^{(n)}$  is used, a suitable value of  $n$  has to be chosen dynamically according to the goodness of  $\hat{V}$ . To aid the manipulation of  $n$ , Sutton (1988) suggested a new estimate constructed by geometrically averaging  $\{V^{(n)}(x) : n \geq 1\}$ .

$$V^\lambda(x) = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} V^{(n)}(x) \quad (25)$$

Here  $(1 - \lambda)$  is a normalizing term. Sutton referred to the learning algorithm that uses  $V^\lambda$  as  $TD(\lambda)$ . Here  $TD$  stands for 'Temporal Difference'. The use of this name will be justified below. Expanding (25) using (22) we get

$$\begin{aligned} V^\lambda(x) &= (1 - \lambda) [V^{(1)}(x) + \lambda V^{(2)}(x) + \lambda^2 V^{(3)}(x) + \dots] \\ &= r_0 + \gamma(1 - \lambda)\hat{V}(x_1; v) + \\ &\quad \gamma\lambda [r_1 + \gamma(1 - \lambda)\hat{V}(x_2; v) + \\ &\quad \gamma\lambda [r_2 + \gamma(1 - \lambda)\hat{V}(x_3; v) + \\ &\quad \dots] \end{aligned} \quad (26)$$

Using the fact that  $r_0 = r(x, \pi(x))$  the above expression may be rewritten recursively as

$$V^\lambda(x) = r(x, \pi(x)) + \gamma(1 - \lambda)\hat{V}(x_1; v) + \gamma\lambda V^\lambda(x_1) \quad (27)$$

where  $x_1$  is the state occurring a time step after  $x$ . Putting  $\lambda = 0$  gives  $V^0 = V^{(1)}$  and putting  $\lambda = 1$  gives  $V^1 = V$ , which is the same as  $V^{(\infty)}$ . Thus, the range of values obtained using  $V^{(n)}$  and varying  $n$  from 1 to  $\infty$  is approximately achieved by using  $V^\lambda$  and varying  $\lambda$  from 0 to 1. A simple idea is to use  $V^\lambda$  instead of  $V^{(n)}$ , begin the learning process with  $\lambda = 1$ , and reduce  $\lambda$  towards zero as learning progresses and  $\hat{V}$  becomes a better estimate of  $V$ . If  $\lambda$  is properly chosen then a significant betterment of computational efficiency is usually achieved when compared to simply using  $\lambda = 0$  or  $\lambda = 1$  (Sutton 1988). In a recent paper, Sutton & Singh (1994) have developed automatic schemes for doing this assuming that no cycles are present in state trajectories.

The definition of  $V^\lambda$  involves all  $V^{(n)}$ s and so it appears that we have to wait for ever to compute it. However, computations involving  $V^\lambda$  can be nicely rearranged and then suitably approximated to yield a practical algorithm *that is suited for doing learning concurrently with real time system operation*. Consider the learning rule in which we use  $V^\lambda$  instead of  $V^{(n)}$ :

$$\hat{V}(x; v) := \hat{V}(x; v) + \beta [V^\lambda(x) - \hat{V}(x; v)] \quad (28)$$

Define the *temporal difference* operator,  $\Delta$  by

$$\Delta(x) = r(x, \pi(x)) + \gamma\hat{V}(x_1; v) - \hat{V}(x; v) \quad (29)$$

$\Delta(x)$  is the difference of predictions (of  $V^\pi(x)$ ) at two consecutive time steps:  $r(x, \pi(x)) + \gamma\hat{V}(x_1; v)$  is a prediction based on information at  $\tau = 1$ , and  $\hat{V}(x; v)$  is a prediction based on information at  $\tau = 0$ . Hence the name, 'temporal difference'. Note that  $\Delta(x)$  can be easily computed using the experience within a time step. A simple rearrangement of the terms in the second line of (26) yields

$$V^\lambda(x) - \hat{V}(x; v) = \Delta(x) + (\gamma\lambda)\Delta(x_1) + (\gamma\lambda)^2\Delta(x_2) + \dots \quad (30)$$

Even (30) is not in a form suitable for use in (28) because it involves future terms,  $\Delta(x_1)$ ,  $\Delta(x_2)$ , etc., extending to infinite time. One way to handle this problem is to choose a large  $N$ , accumulate  $\Delta(x)$ ,  $\Delta(x_1)$ ,  $\dots$ ,  $\Delta(x_{N-1})$  in memory, truncate the right hand side of (30) to include only the first  $N$  terms, and apply (28) at  $\tau = N+1$ , i.e.,  $(N+1)$  time steps after  $x$  occurred. However, a simpler and approximate way of achieving (30) is to include the effects of the temporal differences as and when they occur in time. Let us say that the system is in state  $x$  at time  $t$ . When the system transits to state  $x_1$  at time  $(t+1)$ , compute  $\Delta(x)$  and update  $\hat{V}$  according to:  $\hat{V}(x; v) := \hat{V}(x; v) + \beta(\gamma\lambda)\Delta(x_1)$ . When the system transits to state  $x_2$  at time  $(t+2)$ , compute  $\Delta(x_1)$  and update  $\hat{V}$  according to:  $\hat{V}(x; v) := \hat{V}(x; v) + \beta(\gamma\lambda)^2\Delta(x_2)$  and so on. The reason why this is approximate is because  $\hat{V}(x; v)$  is continuously altered in this process whereas (30) uses the  $\hat{V}(x; v)$  existing at time  $t$ . However, if  $\beta$  is small and so  $\hat{V}(x; v)$  is adapted slowly, the approximate updating method is expected to be close to (28).

One way of implementing the above idea is to maintain an *eligibility trace*,  $e(x, t)$ , for each state visited (Klopf 1972; Klopf 1982; Klopf 1988; Barto *et al* 1983; Watkins 1989), and use the following learning rule at time  $t$ :

$$\hat{V}(x; v) := \hat{V}(x; v) + \beta e(x, t) \Delta(x_t) \quad \forall x \quad (31)$$

where  $x_t$  is the system state at time  $t$ . The eligibility traces can be adapted according to

$$e(x, t) = \begin{cases} 0 & \text{if } x \text{ has never been visited} \\ \gamma\lambda e(x, t-1) & \text{if } x_t \neq x \\ 1 + \gamma\lambda e(x, t-1) & \text{if } x_t = x \end{cases} \quad (32)$$

Two important remarks must be made regarding this implementation scheme.

- Whereas the previous learning rules (e.g., (21), (24) and (28)) update  $\hat{V}$  only for one  $x$  at a time step, (31) updates the  $\hat{V}$  of all states with positive eligibility trace, at a time step. Rule (31) is suitable for connectionist implementation, but not so for implementations on sequential computers. A more efficient way is to keep track of the last  $k$  states visited and update  $\hat{V}$  for them only. The value of  $k$  should depend on  $\lambda$ . If  $\lambda$  is small,  $k$  should be small. If  $\lambda = 0$  then  $k = 1$ .
- The rule for updating eligibility traces, (32) assumes that learning takes place in a single trial. If learning is done over multiple trials then all eligibility traces must be reset to zero just before each new trial is begun.

The remark made below equation (7) applies as well to the learning rules, (21), (24), (28) and, (31). Dayan & Sejnowski (1993), and Jaakkola *et al* (1994) have shown that, if the real time  $TD(\lambda)$  learning rule, (31) is used, then under appropriate assumptions on the variation of  $\beta$  in time, as  $t \rightarrow \infty$ ,  $\hat{V}$  converges to  $V^*$  with probability one. Practically, learning can be achieved by doing multiple trials and decreasing  $\beta$  towards zero as learning progresses.

Thus far in this section we have assumed that the policy,  $\pi$  is deterministic. If  $\pi$  is a stochastic policy then all the ideas of this section still hold with appropriate interpretations: all expectations should include the stochasticity of  $\pi$ , and, the  $\pi(x)$  used in (27), (29) etc. should be taken as instances generated by the stochastic policy.

Let us now come to the estimation of  $Q^\pi$ . Recall from (14) that  $Q^\pi(x, a)$  denotes the total reward obtained by choosing  $a$  as the first action and then following  $\pi$  for all future time steps. Details concerning the extension of  $Q^\pi$  are clearly described in a recent report by Rummery & Niranjan (1994). Let  $\hat{Q}(x, a; v)$  be the estimator of  $Q^\pi(x, a)$  that is to be learnt concurrently with real time system operation. Following the same lines of argument as used for the value function, we obtain a learning rule similar to (31):

$$\hat{Q}(x, a; v) := \hat{Q}(x, a; v) + \beta e_Q(x, a, t) \Delta_Q(x_t, a_t) \quad \forall (x, a) \quad (33)$$

where:  $x_t$  and  $a_t$  are, respectively, the system state and the action chosen at time  $t$ ;

$$\Delta_Q(x, a) = r(x, a) + \gamma \hat{Q}(x_1, \pi(x_1); v) - \hat{Q}(x, a; v); \quad (34)$$

and

$$e_Q(x, a, t) = \begin{cases} 0 & \text{if } (x, a) \text{ has never been visited} \\ \gamma \lambda e_Q(x, a, t-1) & \text{if } (x_t, a_t) \neq (x, a) \\ 1 + \gamma \lambda e_Q(x, a, t-1) & \text{if } (x_t, a_t) = (x, a) \end{cases} \quad (35)$$

As with  $e$ , all  $e_Q(x, a, t)$ 's must be reset to zero whenever a new trial is begun from a random starting state.

If  $\pi$  is a stochastic policy then it is better to replace (34) by

$$\Delta_Q(x, a) = r(x, a) + \gamma \tilde{V}(x_1) - \hat{Q}(x, a; v) \quad (36)$$

where

$$\tilde{V}(x_1) = \sum_{b \in A(x_1)} \text{Prob}\{\pi(x) = b\} \hat{Q}(x_1, b; v) \quad (37)$$

Rummery & Niranjan (1994) suggest the use of (34) even if  $\pi$  is stochastic; in that case, the  $\pi(x_1)$  in (34) corresponds to an instance generated by the stochastic policy at  $x_1$ . We feel that, as an estimate of  $V^\pi(x_1)$ ,  $\tilde{V}(x_1)$  is better than the term  $\hat{Q}(x_1, \pi(x_1); v)$  used in (34), and so it fits in better with the definition of  $Q^\pi$  in (14). Also, if the size of  $A(x_1)$  is small then the computations of  $\tilde{V}(x_1)$  is not much more expensive than that of  $\hat{Q}(x_1, \pi(x_1); v)$ .

## 5. Delayed reinforcement learning methods

Dynamic Programming (DP) methods (Ross 1983; Bertsekas 1989) are well known classical tools for solving the stochastic optimal control problem formulated in §3. Since delayed RL methods also solve the same problem, how do they differ from DP methods?<sup>9</sup> Following are the main differences.

Whereas DP methods simply aim to obtain the optimal value function and an optimal policy using off-line iterative methods, delayed RL methods aim to *learn the same concurrently with real time system operation* and improve performance over time.

<sup>9</sup>The connection between DP and delayed RL was first established by Werbos (1987, 1989, 1992) and Watkins (1989).

- DP methods deal with the complete state space,  $X$  in their computations, while delayed RL methods operate on  $\tilde{X}$ , the set of states that occur during real time system operation. In many applications  $X$  is very large, but  $\tilde{X}$  is only a small, manageable subset of  $X$ . Therefore, in such applications, DP methods suffer from the *curse of dimensionality*, but delayed RL methods do not have this problem. Also, typically delayed RL methods employ function approximators (for value function, policy etc.) that generalize well, and so, after learning, they provide near optimal performance even on unseen parts of the state space.
- DP methods fundamentally require a system model. On the other hand, the main delayed RL methods are model-free; hence they are particularly suited for the on-line learning control of complicated systems for which a model is difficult to derive.
- Because delayed RL methods continuously learn in time they are better suited than DP methods for adapting to situations in which the system and goals are non-stationary.

Although we have said that delayed RL methods enjoy certain key advantages, we should also add that DP has been the fore-runner from which delayed RL methods obtained clues. In fact, it is correct to say that delayed RL methods are basically rearrangements of the computational steps of DP methods so that they can be applied during real time system operation.

Delayed RL methods can be grouped into two categories: model-based methods and model-free methods. Model based methods have direct links with DP. Model-free methods can be viewed as appropriate modifications of the model based methods so as to avoid the model requirement. These methods will be described in detail below.

## 5.1 Model based methods

In this subsection we discuss DP methods and their possible modification to yield delayed RL methods. There are two popular DP methods: value iteration and policy iteration. Value iteration easily extends to give a delayed RL method called 'real time DP'. Policy iteration, though it does not directly yield a delayed method, it forms the basis of an important model-free delayed RL method called actor-critic.

### 5.1.1 Value iteration

The basic idea in value iteration is to compute  $V^*(x)$  as

$$V^*(x) = \lim_{n \rightarrow \infty} V_n^*(x) \quad (38)$$

where  $V_n^*(x)$  is the optimal value function over a finite-horizon of length  $n$ , i.e.,  $V_n^*(x)$  is the maximum expected return if the decision task is terminated  $n$  steps after starting in state  $x$ . For  $n = 1$ , the maximum expected return is just the maximum of the expected immediate payoff:

$$V_1^*(x) = \max_{a \in A(x)} r(x, a) \quad \forall x \quad (39)$$

Then, the recursion,<sup>10</sup>

$$V_{n+1}^*(x) = \max_{a \in A(x)} \left[ r(x, a) + \gamma \sum_y P_{xy}(a) V_n^*(y) \right] \quad \forall x \quad (40)$$

can be used to compute  $V_{n+1}^*$  for  $n = 1, 2, \dots$ . (Iterations can be terminated after a large number ( $N$ ) of iterations, and  $V_N^*$  can be taken to be a good approximation of  $V^*$ .)

In value iteration, a policy is not involved. But it is easy to attach a suitable policy with a value function as follows. Associated with each value function,  $V : X \rightarrow R$  is a policy,  $\pi$  that is *greedy with respect to  $V$* , i.e.,

$$\pi(x) = \arg \max_{a \in A(x)} \left[ r(x, a) + \gamma \sum_y P_{xy}(a) V(y) \right] \quad \forall x \quad (41)$$

If the state space,  $X$  has a very large size (e.g.,  $\text{size} = k^d$ , where  $d$  = number of components of  $x$ ,  $k$  = number of values that each component can take,  $d \approx 10$ ,  $k \approx 100$ ) then value iteration is prohibitively expensive. This difficulty is usually referred to as the *curse of dimensionality*.

In the above, we have assumed that (38) is correct. Let us now prove this convergence. It turns out that convergence can be established for a more general algorithm, of which value iteration is a special case. We call this algorithm as *generalized value iteration*.

### Generalized value iteration

Set  $n = 1$  and  $V_1^*$  = an arbitrary function over states.

Repeat

1. Choose a subset of states,  $B_n$  and set

$$V_{n+1}^*(x) = \begin{cases} \max_{a \in A(x)} \left[ r(x, a) + \gamma \sum_y P_{xy}(a) V_n^*(y) \right] & \text{if } x \in B_n \\ V_n^*(x) & \text{otherwise} \end{cases} \quad (42)$$

2. Reset  $n := n + 1$ .

If we choose  $V_1^*$  as in (39) and take  $B_n = X$  for all  $n$ , then the above algorithm reduces to value iteration. Later we will go into other useful cases of generalized value iteration. But first, let us concern ourselves with the issue of convergence. If  $x \in B_n$ , we will say that the value of state  $x$  has been backed up at the  $n$ -th iteration. Proof of convergence is based on the following result (Bertsekas & Tsitsiklis 1989; Watkins 1989; Barto *et al* 1992).

### Local value improvement theorem

Let  $M_n = \max_x |V_n^*(x) - V^*(x)|$ . Then  $\max_{x \in B_n} |V_{n+1}^*(x) - V^*(x)| \leq \gamma M_n$ .

**Proof:** Take any  $x \in B_n$ . Let  $a^* = \pi^*(x)$  and  $a_n^* = \pi_n^*(x)$ , where  $\pi_n^*$  is a policy that is greedy with respect to  $V_n^*$ . Then

$$\begin{aligned} V_{n+1}^*(x) &\geq r(x, a^*) + \gamma \sum_y P_{xy}(a^*) V_n^*(y) \\ &\geq r(x, a^*) + \gamma \sum_y P_{xy}(a^*) [V^*(y) - M] \\ &= V^*(x) - \gamma M_n \end{aligned}$$

<sup>10</sup>One can also view the recursion as doing a fixed-point iteration to solve Bellman's optimality equation, (13).

Similarly,

$$\begin{aligned} V_{n+1}^*(x) &= r(x, a_n^*) + \gamma \sum_y P_{xy}(a_n^*) V_n^*(y) \\ &\leq r(x, a_n^*) + \gamma \sum_y P_{xy}(a_n^*) [V^*(y) + M] \\ &= V^*(x) + \gamma M_n \end{aligned}$$

and so the theorem is proved.  $\square$

The theorem implies that  $M_{n+1} \leq M_n$  where  $M_{n+1} = \max_x |V_{n+1}^*(x) - V^*(x)|$ . A little further thought shows that the following is also true. If, at the end of iteration  $k$ ,  $K$  further iterations are done in such a way that the value of each state is backed up at least once in these  $K$  iterations, i.e.,  $\cup_{n=k+1}^{k+K} B_n = X$ , then we get  $M_{k+K} \leq \gamma M_k$ . Therefore, if the value of each state is backed up infinitely often, then (38) holds.<sup>11</sup> In the case of value iteration, the value of each state is backed up at each iteration and so (38) holds.

Generalized value iteration was proposed by Bertsekas (1982, 1989) and developed by Bertsekas & Tsitsiklis (1989) as a suitable method of solving stochastic optimal control problems on multi-processor systems with communication time delays and without a common clock. If  $N$  processors are available, the state space can be partitioned into  $N$  sets – one for each processor. The times at which each processor backs up the values of its states can be different for each processor. To back up the values of its states, a processor uses the “present” values of other states communicated to it by other processors.

Barto, Bradtke & Singh (1992) suggested the use of generalized value iteration as a way of learning during real time system operation. They called their algorithm as *Real Time Dynamic Programming* (RTDP). In generalized value iteration as specialized to RTDP,  $n$  denotes system time. At time step  $n$ , let us say that the system resides in state  $x_n$ . Since  $V_n^*$  is available,  $a_n$  is chosen to be an action that is greedy with respect to  $V_n^*$ , i.e.,  $a_n = \pi_n^*(x_n)$ .  $B_n$ , the set of states whose values are backed up, is chosen to include  $x_n$  and, perhaps some more states. In order to improve performance in the immediate future, one can do a lookahead search to some fixed search depth (either exhaustively or by following policy,  $\pi_n^*$ ) and include these probable future states in  $B_n$ . Because the value of  $x_n$  is going to undergo change at the present time step, it is a good idea to also include, in  $B_n$ , the most likely predecessors of  $x_n$  (Moore & Atkeson 1993).

One may ask: since a model of the system is available, why not simply do value iteration or, do generalized value iteration as Bertsekas & Tsitsiklis suggest? In other words, what is the motivation behind RTDP? The answer is simple. In most problems (e.g., playing games such as checkers and backgammon) the state space is extremely large, but only a small subset of it actually occurs during usage. Because RTDP works concurrently with actual system operation, it focusses on regions of the state space that are most relevant to the system's behaviour. For instance, successful learning was accomplished in the checkers program of Samuel (1959) and in the backgammon program, TDgammon of Tesauro (1992) using variations of RTDP. In (Barto *et al* 1992), Barto, Bradtke & Singh also use RTDP to make interesting connections and useful extensions to learning real time search algorithms in Artificial Intelligence (Korf 1990).

The convergence result mentioned earlier says that the values of all states have to

<sup>11</sup> If  $\gamma = 1$ , then convergence holds under certain assumptions. The analysis required is more sophisticated. See (Bertsekas & Tsitsiklis 1989; Bradtke 1994) for details.



be backed up infinitely often<sup>12</sup> in order to ensure convergence. So it is important to suitably explore the state space in order to improve performance. Barto, Bradtke & Singh have suggested two ways of doing exploration<sup>13</sup>: (1) adding stochasticity to the policy; and (2) doing learning cumulatively over multiple trials.

If, only an inaccurate system model is available then it can be updated in real time using a system identification technique, such as maximum likelihood estimation method (Barto *et al* 1992). The current system model can be used to perform the computations in (42). Convergence of such adaptive methods has been proved by Gullapalli & Barto (1994).

### 5.1.2 Policy iteration

Policy iteration operates by maintaining a representation of a policy and its value function, and forming an improved policy using them. Suppose  $\pi$  is a given policy and  $V^\pi$  is known. How can we improve  $\pi$ ? An answer will become obvious if we first answer the following simpler question. If  $\mu$  is another given policy then when is

$$V^\mu(x) \geq V^\pi(x) \quad \forall x \quad (43)$$

i.e., when is  $\mu$  uniformly better than  $\pi$ ? The following simple theorem (Watkins 1989) gives the answer.

#### Policy improvement theorem

The policy  $\mu$  is uniformly better than policy  $\pi$  if

$$Q^\pi(x, \mu(x)) \geq V^\pi(x) \quad \forall x \quad (44)$$

**Proof:** To avoid clumsy details let us give a not-so-rigorous proof (Watkins 1989). Starting at  $x$ , it is better to follow  $\mu$  for one step and then to follow  $\pi$ , than it is to follow  $\pi$  right from the beginning. By the same argument, it is better to follow  $\mu$  for one further step from the state just reached. Repeating the argument we get that it is always better to follow  $\mu$  than  $\pi$ . See Bellman & Dreyfus (1962) and Ross (1983) for a detailed proof.  $\square$

Let us now return to our original question: given a policy  $\pi$  and its value function  $V^\pi$ , how do we form an improved policy,  $\mu$ ? If we define  $\mu$  by

$$\mu(x) = \arg \max_{a \in A(x)} Q^\pi(x, a) \quad \forall x \quad (45)$$

then (44) holds. By the policy improvement theorem  $\mu$  is uniformly better than  $\pi$ . This is the main idea behind policy iteration.

#### Policy iteration

Set  $\pi :=$  an arbitrary initial policy and compute  $V^\pi$ .

Repeat

1. Compute  $Q^\pi$  using (14).
2. Find  $\mu$  using (45) and compute  $V^\mu$ .

<sup>12</sup>For good practical performance it is sufficient that states that are most relevant to the system's behaviour are backed up repeatedly.

<sup>13</sup>Thrun (1986) has discussed the importance of exploration and suggested a variety of methods for it

3. Set:  $\pi := \mu$  and  $V^\pi := V^\mu$ .

until  $V^\mu = V^\pi$  occurs at step 2.

Nice features of the above algorithm are: (1) it terminates after a finite number of iterations because there are only a finite number of policies; and (2) when termination occurs we get

$$V^\pi(x) = \max_a Q^\pi(x, a) \quad \forall x$$

(i.e.,  $V^\pi$  satisfies Bellman's optimality equation) and so  $\pi$  is an optimal policy. But the algorithm suffers from a serious drawback: it is very expensive because the entire value function associated with a policy has to be recalculated at each iteration (step 2). Even though  $V^\mu$  may be close to  $V^\pi$ , unfortunately there is no simple short cut to compute it. In §5.2 we will discuss a well-known model-free method called the *actor-critic* method which gives an inexpensive approximate way of implementing policy iteration.

## 5.2 Model-free methods

Model-free delayed RL methods are derived by making suitable approximations to the computations in value iteration and policy iteration, so as to eliminate the need for a system model. Two important methods result from such approximations: Barto, Sutton & Anderson's actor-critic (Barto *et al* 1983), and Watkins' Q-Learning (Watkins 1989). These methods are milestone contributions to the optimal feedback control of dynamic systems.

### 5.2.1 Actor-critic method

The actor-critic method was proposed by Barto *et al* (1983) (in their popular work on balancing a pole on a moving cart) as a way of combining, on a step-by-step basis, the process of forming the value function with the process of forming a new policy. The method can also be viewed as a practical, approximate way of doing policy iteration: perform one step of an on-line procedure for estimating the value function for a given policy, and at the same time perform one step of an on-line procedure for improving that policy. The actor-critic method<sup>14</sup> is best derived by combining the ideas of §2 and §4 on immediate RL and estimating value function, respectively. Details are as follows.

**Actor ( $\pi$ )** Let  $m$  denote the total number of actions. Maintain an approximator,  $g(\cdot; w) : X \rightarrow R^m$  so that  $z = g(x; w)$  is a vector of merits of the various feasible actions at state  $x$ . In order to do exploration, choose actions according to a stochastic action selector such as (6).<sup>15</sup>

**Critic ( $V^\pi$ )** Maintain an approximator,  $\hat{V}(\cdot; w) : X \rightarrow R$  that estimates the value function (expected total reward) corresponding to the stochastic policy mentioned above. The ideas of §4 can be used to update  $\hat{V}$ .

Let us now consider the process of learning the actor. Unlike immediate RL, learning is more complicated here for the following reason. Whereas, in immediate

<sup>14</sup> A mathematical analysis of this method has been done by Williams & Baird (1990).

<sup>15</sup> In their original work on pole-balancing, Barto, Sutton & Anderson suggested a different way of including stochasticity.

RL the environment immediately provides an evaluation of an action, in delayed RL the effect of an action on the total reward is not immediately available and has to be estimated appropriately. Suppose, at some time step, the system is in state  $x$  and the action selector chooses action  $a^k$ . For  $g$  the learning rule that parallels (5) would be

$$g_k(x; w) := g_k(x; w) + \alpha [\rho(x, a^k) - \hat{V}(x; v)] \quad (46)$$

where  $\rho(x; a^k)$  is the expected total reward obtained if  $a^k$  is applied to the system at state  $x$  and then policy  $\pi$  is followed from the next step onwards. An approximation is

$$\rho(x, a^k) \approx r(x, a^k) + \gamma \sum_y P_{xy}(a^k) \hat{V}(y; v) \quad (47)$$

This estimate is unavailable because we do not have a model. A further approximation is

$$\rho(x, a^k) \approx r(x, a^k) + \gamma \hat{V}(x_1; v) \quad (48)$$

where  $x_1$  is the state occurring in the real time operation when action  $a^k$  is applied at state  $x$ . Using (48) in (46) gives

$$g_k(x; w) := g_k(x; w) + \alpha \Delta(x) \quad (49)$$

where  $\Delta$  is as defined in (29). The following algorithm results.

#### Actor-critic trial

Set  $t = 0$  and  $x = a$  random starting state.

Repeat (for a number of time steps)

1. With the system at state,  $x$ , choose action  $a$  according to (6) and apply it to the system. Let  $x_1$  be the resulting next state.
2. Compute  $\Delta(x) = r(x, a) + \gamma \hat{V}(x_1; v) - \hat{V}(x; v)$
3. Update  $\hat{V}$  using  $\hat{V}(x; v) := \hat{V}(x; v) + \beta \Delta(x)$
4. Update  $g_k$  using (49) where  $k$  is such that  $a = a^k$ .

The above algorithm uses the  $TD(0)$  estimate of  $V^\pi$ . To speed-up learning the  $TD(\lambda)$  rule, (31) can be employed. Barto *et al* (1983) and others (Gullapalli 1992a; Gullapalli *et al* 1994) use the idea of eligibility traces for updating  $g$  also. They give only an intuitive explanation for this usage. Lin (1992) has suggested the accumulation of data until a trial is over, update  $\hat{V}$  using (28) for all states visited in the trial, and then update  $g$  using (49) for all (state, action) pairs experienced in the trial.

### 5.2.2 Q-Learning

Just as the actor-critic method is a model-free, on-line way of approximately implementing policy iteration, Watkins' Q-Learning algorithm is a model-free, on-line way of approximately implementing generalized value iteration. Though the RTDP algorithm does generalized value iteration concurrently with real time system operation, it requires the system model for doing a crucial operation: the determination of the maximum on the right hand side of (42). Q-Learning overcomes this problem

elegantly by operating with the  $Q$ -function instead of the value function. (Recall, from §3, the definition of  $Q$ -function and the comment on its advantage over value function.)

The aim of  $Q$ -Learning is to find a function approximator,  $\hat{Q}(\cdot, \cdot; v)$  that approximates  $Q^*$ , the solution of Bellman's optimality equation, (16), in on-line mode without employing a model. However, for the sake of developing ideas systematically, let us begin by assuming that a system model is available and consider the modification of the ideas of §5.1a to use the  $Q$ -function instead of the value function. If we think in terms of a function approximator,  $\hat{V}(x; v)$  for the value function, the basic update rule that is used throughout §5.1a is

$$\hat{V}(x; v) := \max_{a \in A(x)} \left[ r(x, a) + \gamma \sum_y P_{xy}(a) \hat{V}(y; v) \right]$$

For the  $Q$ -function, the corresponding rule is

$$\hat{Q}(x, a; v) := r(x, a) + \gamma \sum_y P_{xy}(a) \max_{b \in A(y)} \hat{Q}(y, b; v) \quad (50)$$

Using this rule, all the ideas of §5.1a can be easily modified to employ the  $Q$ -function.

However, our main concern is to derive an algorithm that avoids the use of a system model. A model can be avoided if we: (1) replace the summation term in (50) by  $\max_{b \in A(x_1)} \hat{Q}(x_1, b; v)$  where  $x_1$  is an instance of the state resulting from the application of action  $a$  at state  $x$ ; and (2) achieve the effect of the update rule in (50) via the "averaging" learning rule,

$$\hat{Q}(x, a; v) := \hat{Q}(x, a; v) + \beta \left[ r(x, a) + \gamma \max_{b \in A(x_1)} \hat{Q}(x_1, b; v) - \hat{Q}(x, a; v) \right] \quad (51)$$

If (51) is carried out we say that the  $Q$ -value of  $(x, a)$  has been backed up. Using (51) in on-line mode of system operation we obtain the  $Q$ -Learning algorithm.

#### $Q$ -Learning trial

Set  $t = 0$  and  $x = a$  random starting state.

Repeat (for a number of time steps)

1. Choose action  $a \in A(x)$  and apply it to the system. Let  $x_1$  be the resulting state.
2. Update  $\hat{Q}$  using (51).
3. Reset  $x := y$ .

The remark made below equation, (7) in §2 is very appropriate for the learning rule, (51). Watkins showed<sup>16</sup> that if the  $Q$ -value of each admissible  $(x, a)$  pair is backed up infinitely often, and if the step size,  $\beta$  is decreased to zero in a suitable way then as  $t \rightarrow \infty$ ,  $\hat{Q}$  converges to  $Q^*$  with probability one. Practically, learning can be achieved by: (1) using, in step 1, an appropriate exploration policy that tries all

<sup>16</sup>A revised proof was given by Watkins & Dayan (1992). Tsitsiklis (1993) and Jaakkola et al., (1994) have given other proofs.

actions;<sup>17</sup> (2) doing multiple trials to ensure that all states are frequently visited; and (3) decreasing  $\beta$  towards zero as learning progresses.

We now discuss a way of speeding up  $Q$ -Learning by using the  $TD(\lambda)$  estimate of the  $Q$ -function, derived in §4. If  $TD(\lambda)$  is to be employed in a  $Q$ -Learning trial, a fundamental requirement is that the policy used in step 1 of the  $Q$ -Learning Trial and the policy used in the update rule, (51) should match (note the use of  $\pi$  in (34) and (37)). Thus  $TD(\lambda)$  can be used if we employ the greedy policy,

$$\pi(x) = \arg \max_{a \in A(x)} \hat{Q}(x, a; v) \quad (52)$$

in step 1.<sup>18</sup> But, this leads to a problem: use of the greedy policy will not allow exploration of the action space, and hence poor learning can occur. Rummery & Niranjan (1994) give a nice comparative account of various attempts described in the literature for dealing with this conflict. Here we only give the details of an approach that Rummery & Niranjan found to be very promising.

Consider the stochastic policy (based on the Boltzmann distribution and  $Q$ -values),

$$\text{Prob}\{\pi(x) = a|x\} = \frac{\exp(\hat{Q}(x, a; v)/T)}{\sum_{b \in A(x)} \exp(\hat{Q}(x, b; v)/T)}, \quad a \in A(x) \quad (53)$$

where  $T \in [0, \infty)$ . When  $T \rightarrow \infty$  all actions have equal probabilities and, when  $T \rightarrow 0$  the stochastic policy tends towards the greedy policy in (52). To learn,  $T$  is started with a suitable large value (depending on the initial size of the  $Q$ -values) and is decreased to zero using an annealing rate; at each  $T$  thus generated, multiple  $Q$ -learning trials are performed. This way, exploration takes place at the initial large  $T$  values. The  $TD(\lambda)$  learning rule, (36) estimates expected returns for the policy at each  $T$ , and, as  $T \rightarrow 0$ ,  $\hat{Q}$  will converge to  $Q^*$ . The ideas here are somewhat similar to those of simulated annealing.

### 5.3 Extension to continuous spaces

Optimal control of dynamic systems typically involves the solution of delayed RL problems having continuous state/action spaces. If the state space is continuous but the action space is discrete then all the delayed RL algorithms discussed earlier can be easily extended, provided appropriate function approximators that generalize a real time experience at a state to all topologically nearby states are used; see §6 for a discussion of such approximators. On the other hand, if the action space is continuous, extension of the algorithms is more difficult. The main cause of the difficulty can be easily seen if we try extending RTDP to continuous action spaces: the max operation in (42) is non-trivial and difficult if  $A(x)$  is continuous. (Therefore, even methods based on value iteration need to maintain a function approximator for actions.) In the rest of this subsection we will give a brief review of

<sup>17</sup>Note that step 1 does not put any restriction on choosing a feasible action. So, any stochastic exploration policy that, at every  $x$  generates each feasible action with positive probability can be used. When learning is complete, the greedy policy,  $\pi(x) = \arg \max_{a \in A(x)} \hat{Q}(x, a; v)$  should be used for optimal system performance.

<sup>18</sup>If more than one action attains the maximum in (52) then for convenience we take  $\pi$  to be a stochastic policy that makes all such maximizing actions equally probable.

various methods of handling continuous action spaces. Just to make the presentation easy, we will make the following assumptions.

- The system being controlled is deterministic. Let

$$x_{t+1} = f(x_t, a_t) \quad (54)$$

describe the transition.<sup>19</sup>

- There are no action constraints, i.e.,  $A(x)$  = an  $m$ -dimensional real space for every  $x$ .
- All functions involved ( $r, f, \hat{V}, \hat{Q}$  etc.) are continuously differentiable.

Let us first consider model-based methods. Werbos (1990b) has proposed a variety of algorithms. Here we will describe only one important algorithm, the one that Werbos refers to as *Backpropagated Adaptive Critic*. The algorithm is of the actor-critic type, but it is somewhat different from the actor-critic method of §5.2a. There are two function approximators:  $\hat{\pi}(\cdot; w)$  for action; and,  $\hat{V}(\cdot; v)$  for critic. The critic is meant to approximate  $V^\pi$ ; at each time step, it is updated using the  $TD(\lambda)$  learning rule, (31) of §4. The actor tries to improve the policy at each time step using the hint provided by the policy improvement theorem in (44). To be more specific, let us define

$$Q(x, a) \stackrel{\text{def}}{=} r(x, a) + \gamma \hat{V}(f(x, a); v) \quad (55)$$

At time  $t$ , when the system is at state  $x_t$ , we choose the action,  $a_t = \hat{\pi}(x_t; w)$ , leading to the next state,  $x_{t+1}$  given by (54). Let us assume  $\hat{V} = V^\pi$ , so that  $V^\pi(x_t) = Q(x_t, a_t)$  holds. Using the hint from (44), we aim to adjust  $\hat{\pi}(x_t; w)$  to give a new value,  $a^{\text{new}}$  such that

$$Q(x_t, a^{\text{new}}) > Q(x_t, a_t) \quad (56)$$

A simple learning rule that achieves this requirement is

$$\hat{\pi}(x_t; w) := \hat{\pi}(x_t; w) + \alpha \frac{\partial Q(x_t, a)}{\partial a} \Big|_{a=a_t} \quad (57)$$

where  $\alpha$  is a small (positive) step size. The partial derivative in (57) can be evaluated using

$$\frac{\partial Q(x_t, a)}{\partial a} = \frac{\partial r(x_t, a)}{\partial a} + \gamma \frac{\partial \hat{V}(y; v)}{\partial y} \Big|_{y=f(x_t, a)} \frac{\partial f(x_t, a)}{\partial a} \quad (58)$$

Let us now come to model-free methods. A simple idea is to adapt a function approximator,  $\hat{f}$  for the system model function,  $f$ , and use  $\hat{f}$  instead of  $f$  in Werbos' algorithm. On-line experience, i.e., the combination,  $(x_t, a_t, x_{t+1})$ , can be used to learn  $\hat{f}$ . This method was proposed by Brody (1992), actually as a way of overcoming a serious deficiency<sup>20</sup> associated with an ill-formed model-free method suggested by Jordan & Jacobs (1990). A key difficulty associated with Brody's method is that, until the learning system adapts a good  $\hat{f}$ , system performance does not improve

<sup>19</sup> Werbos (1990b) describes ways of treating stochastic systems.

<sup>20</sup> This deficiency was also pointed out by Gullapalli (1992b).

at all; in fact, at the early stages of learning the method can perform in a confused way. To overcome this problem Brody suggests that  $\hat{f}$  be learnt well, before it is used to train the actor and the critic.

A more direct model-free method can be derived using the ideas of §5.2a and employing a learning rule similar to (8) for adapting  $\hat{\pi}$ . This method was proposed and successfully demonstrated by Gullapalli (Gullapalli 1992a; Gullapalli *et al* 1994). Since Gullapalli's method learns by observing the effect of a randomly chosen perturbation of the policy, it is not as systematic as the policy change in Brody's method.

We now propose a new model-free method that systematically changes the policy similar to what Brody's method does, and, avoids the need for adapting a system model. This is achieved using a function approximator,  $\hat{Q}(\cdot, \cdot; v)$  for approximating  $Q^{\pi}$ , the  $Q$ -function associated with the actor. The  $TD(\lambda)$  learning rule in (33) can be used for updating  $\hat{Q}$ . Also, policy improvement can be attempted using the learning rule (similar to (57)),

$$\hat{\pi}(x_t; w) := \hat{\pi}(x_t; w) + \alpha \frac{\partial \hat{Q}(x_t, a)}{\partial a} \Big|_{a=a_t} \quad (59)$$

We are currently performing simulations to study the performance of this new method relative to the other two model-free methods mentioned above.

Werbos' algorithm and our  $Q$ -Learning based algorithm are deterministic, while Gullapalli's algorithm is stochastic. The deterministic methods are expected to be much faster, whereas the stochastic method has better assurance of convergence to the true solution. The arguments are similar to those mentioned at the end of §2.

## 6. Other issues

In this section we discuss a number of issues relevant to practical implementation of RL algorithms. A nice discussion of these (and other) issues has been presented by Barto (1992).

### 6.1 Function-approximation

A variety of function approximators has been employed by researchers to practically solve RL problems. When the input space of the function approximator is finite, the most straight-forward method is to use a *look-up table* (Singh 1992a; Moore & Atkeson 1993). All theoretical results on the convergence of RL algorithms assume this representation. The disadvantage of using look-up table is that if the input space is large then the memory requirement becomes prohibitive.<sup>21</sup> Continuous input spaces have to be discretized when using a look-up table. If the discretization is done finely so as to obtain good accuracy we have to face the 'curse of dimensionality'. One way of overcoming this is to do a problem-dependent discretization; see, for example, the 'BOXES' representation used by Barto *et al* (1983) and others (Michie & Chambers 1968; Gullapalli *et al* 1994; Rosenet *et al* 1991) to solve the pole balancing problem.

<sup>21</sup> Buckland & Lawrence (1994) have proposed a new delayed RL method called Transition point DP which can significantly reduce the memory requirement for problems in which optimal actions change infrequently in time.

Non look-up table approaches use parametric function approximation methods. These methods have the advantage of being able to generalize beyond the training data and hence give reasonable performance on unvisited parts of the input space. Among these, connectionist methods are the most popular. Connectionist methods that have been employed for RL can be classified into four groups: multi-layer perceptrons; methods based on clustering; CMAC; and recurrent networks. *Multi-layer perceptrons* have been successfully used by Anderson (1986, 1989) for pole balancing, Lin (1991a, 1991b, 1991c, 1992) for a complex test problem, Tesauro (1992) for backgammon, Thrun (1993) and Millan & Torras (1992) for robot navigation, and others (Boyan 1992; Gullapalliet *al* 1994). On the other hand, Watkins (1989), Chapman (1991), Kaelbling (1990, 1991), and Shepanski & Macy (1987) have reported bad results. A modified form of Platt's *Resource allocation network* (Platt 1991), a method based on radial basis functions, has been used by Anderson (1993) for pole balancing. Many researchers have used CMAC (Albus 1975) for solving RL problems: Watkins (1989) for a test problem; Singh (1991, 1992b, 1992d) and Tham & Prager (1994) for a navigation problem; Lin & Kim (1991) for pole balancing; and Sutton (1990, 1991b) in his 'Dyna' architecture. Recurrent networks with context information feedback have been used by Bacharach (1991, 1992) and Mozer & Bacharach (1990a, 1990b) in dealing with RL problems with incomplete state information.

A few non-connectionist methods have also been used for RL. Mahadevan & Connell (1991) have used statistical clustering in association with *Q*-Learning for the automatic programming of a mobile robot. A novel feature of their approach is that the number of clusters is dynamically varied. Chapman & Kaelbling (1991) have used a tree-based clustering approach in combination with a modified *Q*-Learning algorithm for a difficult test problem with a huge input space.

The function approximator has to exercise care to ensure that learning at some input point,  $x$  does not seriously disturb the function values for  $y \neq x$ . It is often advantageous to choose a function approximator and employ an update rule in such a way that the function values of  $x$  and states 'near'  $x$  are modified similarly while the values of states 'far' from  $x$  are left unchanged. Such a choice usually leads to good generalization, i.e., good performance of the learnt function approximator even on states that are not visited during learning. In this respect, CMAC and methods based on clustering, such as RBF, statistical clustering, etc., are more suitable than multi-layer perceptrons.

The effect of errors introduced by function approximators on the optimal performance of the controller has not been well understood.<sup>22</sup> It has been pointed out by Watkins (1989), Bradtke (1993), and others (Barto 1992), that, if function approximation is not done in a careful way, poor learning can result. In the context of *Q*-Learning, Thrun & Schwartz (1993) have shown that errors in function approximation can lead to a systematic over estimation of the *Q*-function. Linden (1993) points out that in many problems the value function is discontinuous and so using continuous function approximators is inappropriate. But he does not suggest any clear remedies for this problem. Overall, it must be mentioned that much work needs to be done on the use of function approximators for RL.

<sup>22</sup>Bertsekas (1989) and Singh & Yee (1993) have derived some theoretical bounds for errors in value function in terms of function approximator error.



## 6.2 Modular and hierarchical architectures

When applied to problems with large task space or sparse rewards, RL methods are terribly slow to learn. Dividing the problem into simpler subproblems, using a hierarchical control structure, etc., are ways of overcoming this.

*Sequential task decomposition* is one such method. This method is useful when a number of complex tasks can be performed making use of a finite number of “elemental” tasks or skills, say,  $T_1, T_2, \dots, T_n$ . The original objective of the controller can then be achieved by temporally concatenating a number of these elemental tasks to form what is called a “composite” task. For example,

$$C_j = [T(j, 1), T(j, 2), \dots, T(j, k)], \quad \text{where } T(j, i) \in \{T_1, T_2, \dots, T_n\}$$

is a composite task made up of  $k$  elemental tasks that have to be performed in the order listed. Reward functions are defined for each of the elemental tasks, making them more abundant than in the original problem definition.

Singh (1992a, 1992b) has proposed an algorithm based on a modular connectionist network (Jacobson *et al* 1991), making use of these ideas. In his work the controller is unaware of the decomposition of the task and has to learn both the elemental tasks, and the decomposition of the composite tasks simultaneously. Tham & Prager (1994) and Lin (1993) have proposed similar solutions. Mahadevan & Connel (1991) have developed a method based on the *subsumption architecture* (Brooks 1986) where the decomposition of the task is specified by the user before hand, and the controller learns only the elemental tasks, while Maes & Brooks (1990) have shown that the controller can be made to learn the decomposition also, in a similar framework. All these methods require some external agency to specify the problem decomposition. Can the controller itself learn how the problem is to be decomposed? Though Singh (1992d) has some preliminary results, much work needs to be done here.

Another approach to this problem is to use some form of hierarchical control (Watkins 1989). Here there are different “levels” of controllers<sup>23</sup>, each learning to perform a more abstract task than the level below it and directing the lower level controllers to achieve its objective. For example, in a ship a navigator decides in what direction to sail so as to reach the port while the helmsman steers the ship in the direction indicated by the navigator. Here the navigator is the higher level controller and the helmsman the lower level controller. Since the higher level controllers have to work on a smaller task space and the lower level controllers are set simpler tasks improved performance results.

Examples of such hierarchical architectures are *Feudal RL* by Dayan & Hinton (1993) and *Hierarchical planning* by Singh (1992a, 1992c). These methods too, require an external agency to specify the hierarchy to be used. This is done usually by making use of some “structure” in the problem.

Training controllers on simpler tasks first and then training them to perform progressively more complex tasks using these simpler tasks, can also lead to better performance. Here at any one stage the controller is faced with only a simple learning task. This technique is called *shaping* in animal behaviour literature. Gullapalli (1992a) and Singh (1992d) have reported some success in using this idea. Singh shows that the controller can be made to “discover” a decomposition of the task by itself using this technique.

<sup>23</sup> Controllers at different levels may operate at different temporal resolutions.

### 6.3 Speeding-up learning

Apart from the ideas mentioned above, various other techniques have been suggested for speeding-up RL. Two novel ideas have been suggested by Lin (1991a, 1991b, 1991c, 1992): *experience playback*; and *teaching*. Let us first discuss experience playback. An experience consists of a quadruple (occurring in real time system operation),  $(x, a, y, r)$ , where  $x$  is a state,  $a$  is the action applied at state  $x$ ,  $y$  is the resulting state, and  $r$  is  $r(x, a)$ . Past experiences are stored in a finite memory buffer,  $\mathcal{P}$ . An appropriate strategy can be used to maintain  $\mathcal{P}$ . At some point in time let  $\pi$  be the "current" (stochastic) policy. Let

$$\mathcal{E} = \{(x, a, y, r) \in \mathcal{P} \mid \text{Prob}\{\pi(x) = a\} \geq \epsilon\}$$

where  $\epsilon$  is some chosen tolerance. The learning update rule is applied, not only to the current experience, but also to a chosen subset of  $\mathcal{E}$ . Experience playback can be especially useful in learning about rare experiences. In teaching, the user provides the learning system with experiences so as to expedite learning.

Incorporating domain specific knowledge also helps in speeding-up learning. For example, for a given problem, a "nominal" controller that gives reasonable performance may be easily available. In that case RL methods can begin with this controller and improve its performance (Singh *et al* 1994). Domain specific information can also greatly help in choosing state representation and setting up the function approximators (Barto 1992; Millan & Torras 1992).

In many applications an inaccurate system model is available. It turns out to be very inefficient to discard the model and simply employ a model-free method. An efficient approach is to interweave a number of "planning" steps between every two on-line learning steps. A planning step may be one of the following: a time step of a model-based method such as RTDP; or, a time step of a model-free method for which experience is generated using the available system model. In such an approach, it is also appropriate to adapt the system model using on-line experience. These ideas form the basis of Sutton's *Dyna* architectures (Sutton 1990, 1991b) and related methods (Moore & Atkeson 1993; Peng & Williams 1993).

## 7. Conclusion

In this paper we have tried to give a cohesive overview of existing RL algorithms. Though research has reached a mature level, RL has been successfully demonstrated only on a few practical applications (Gullapalli *et al* 1994; Tesauro 1992; Mahadevan & Connell 1991; Thrun 1993), and clear guidelines for its general applicability do not exist. The connection between DP and RL has nicely bridged control theorists and AI researchers. With contributions from both these groups on the pipeline, more interesting results are forthcoming and it is expected that RL will make a strong impact on the intelligent control of dynamic systems.

## References

- Albus J S 1975 A new approach to manipulator control: The cerebellar model articulation controller (CMAC). *Trans. ASME, J. Dyn. Syst., Meas., Contr.*

97:220-227

- Anderson C W 1986 *Learning and problem solving with multilayer connectionist systems*. Ph D thesis, University of Massachusetts, Amherst, MA
- Anderson C W 1987 Strategy learning with multilayer connectionist representations. Technical report, TR87-509.3, GTE Laboratories, INC., Waltham, MA
- Anderson C W 1989 Learning to control an inverted pendulum using neural networks. *IEEE Contr. Syst. Mag.* : 31-37
- Anderson C W 1993 Q-Learning with hidden-unit restarting. In *Advances in neural information processing systems 5* (eds) S J Hanson, J D Cowan, C L Giles (San Mateo, CA: Morgan Kaufmann) pp 81-88
- Bacharach J R 1991 A connectionist learning control architecture for navigation. In *Advances in neural information processing systems 3* (eds) R P Lippman, J E Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 457-463
- Bacharach J R 1992 *Connectionist modeling and control of finite state environments*. Ph D thesis, University of Massachusetts, Amherst, MA
- Barto A G 1985 Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology* 4: 229-256
- Barto A G 1986 Game-theoretic cooperativity in networks of self interested units. In *Neural networks for computing* (ed.) J S Denker (New York: American Institute of Physics) pp 41-46
- Barto A G 1992 Reinforcement learning and adaptive critic methods. In *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches* (eds) D A White, D A Sofge (New York: Van Nostrand Reinhold) pp 469-491
- Barto A G, Anandan P 1985 Pattern recognizing stochastic learning automata. *IEEE Trans. Syst., Man Cybern.* 15: 360-375
- Barto A G, Anandan P, Anderson C W 1985 Cooperativity in networks of pattern recognizing stochastic learning automata. In *Proceedings of the Fourth Yale Workshop on Applications of Adaptive Systems Theory*, New Haven, CT
- Barto A G, Bradtke S J, Singh S P 1992 Real-time learning and control using asynchronous dynamic programming. Technical Report COINS 91-57, University of Massachusetts, Amherst, MA
- Barto A G, Jordan M I 1987 Gradient following without back-propagation in layered networks. In *Proceedings of the IEEE First Annual Conference on Neural Networks*, (eds) M Caudill, C Butler (New York: IEEE) pp II629-II636
- Barto A G, Singh S P 1991 On the computational economics of reinforcement learning. In *Connectionist Models Proceedings of the 1990 Summer School*. (eds) D S Touretzky, J L Elman, T J Sejnowski, G E Hinton (San Mateo, CA: Morgan Kaufmann) pp 35-44

- Barto A G, Sutton R S 1981, Landmark learning: an illustration of associative search. *Biol. Cybern.* 42: 1-8
- Barto A G, Sutton R S 1982 Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behav. Brain Res.* 4: 221-235
- Barto A G, Sutton R S, Anderson C W 1983 Neuronlike elements that can solve difficult learning control problems. *IEEE Trans. Syst., Man Cybern.* 13: 835-846
- Barto A G, Sutton R S, Brouwer P S 1981 Associative search network: a reinforcement learning associative memory. *IEEE Trans. Syst., Man Cybern.* 40: 201-211
- Barto A G, Sutton R S, Watkins C J C H 1990 Learning and sequential decision making. In *Learning and computational neuroscience: Foundations of adaptive networks*. (eds) M Gabriel, J Moore (Cambridge, MA: MIT Press) pp 539-602
- Bellman R E, Dreyfus S E 1962 *Applied dynamic programming*. RAND Corporation
- Bertsekas D P 1982 Distributed dynamic programming. *IEEE Trans. Autom. Contr.* 27: 610-616
- Bertsekas D P 1989 *Dynamic programming: Deterministic and stochastic models* (Englewood Cliffs, NJ: Prentice-Hall)
- Bertsekas D P, Tsitsiklis J N 1989 *Parallel and distributed computation: Numerical methods* (Englewood Cliffs, NJ: Prentice-Hall)
- Boyen J 1992 *Modular neural networks for learning context-dependent game strategies*. Masters thesis, Computer Speech and Language Processing, University of Cambridge, Cambridge, England
- Bradtke S J 1993 Reinforcement learning applied to linear quadratic regulation. In *Advances in neural information processing systems 5* (eds) S J Hanson, J D Cowan, C L Giles (San Mateo, CA: Morgan Kaufmann) pp 295-302
- Bradtke S J 1994 *Incremental dynamic programming for on-line adaptive optimal control*. CMPSCI Technical Report 94-62
- Brody C 1992 Fast learning with predictive forward models. In *Advances in neural information processing systems 4* (eds) J E Moody, S J Hanson, R P Lippmann (San Mateo, CA: Morgan Kaufmann) pp 563-570
- Brooks R A 1986 Achieving artificial intelligence through building robots. Technical Report, A I Memo 899, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, Cambridge, MA
- Buckland K M, Lawrence P D 1994 Transition point dynamic programming. In *Advances in neural information processing systems 6* (eds) J D Cowan, G Tesauro, J Alsppector (San Francisco, CA: Morgan Kaufmann) pp 639-646
- Chapman D 1991 *Vision, Instruction, and Action* (Cambridge, MA: MIT Press)

- Chapman D, Kaelbling L P 1991 Input generalization in delayed reinforcement learning: an algorithm and performance comparisons. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*
- Chrisman L 1992 Planning for closed-loop execution using partially observable markovian decision processes. In *Proceedings of AAAI*
- Dayan P 1991a Navigating through temporal difference. In *Advances in neural information processing systems 3* (eds) R P Lippmann, J E Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 464-470
- Dayan P 1991b *Reinforcing connectionism: Learning the statistical way*. Ph D thesis, University of Edinburgh, Edinburgh
- Dayan P, Hinton G E 1993 Feudal reinforcement learning. In *Advances in neural information processing systems 5* (eds) S J Hanson, J D Cowan, C L Giles (San Mateo, CA: Morgan Kaufmann) pp 271-278
- Dayan P, Sejnowski T J 1993 TD( $\lambda$ ) converges with probability 1. Technical Report, CNL, The Salk Institute, San Diego, CA
- Dean T L, Wellman M P 1991 *Planning and control* (San Mateo, CA: Morgan Kaufmann)
- Gullapalli V 1990 A stochastic reinforcement algorithm for learning real-valued functions. *Neural Networks* 3: 671-692
- Gullapalli V 1992a Reinforcement learning and its application to control. Technical Report, COINS, 92-10, Ph D thesis, University of Massachusetts, Amherst, MA
- Gullapalli V 1992b A comparison of supervised and reinforcement learning methods on a reinforcement learning task. In *Proceedings of the 1991 IEEE Symposium on Intelligent Control*, Arlington, VA
- Gullapalli V, Barto A G 1994 Convergence of indirect adaptive asynchronous value iteration algorithms. In *Advances in neural information processing systems 6* (eds) J D Cowan, G Tesauro, J Alsppector (San Francisco, CA: Morgan Kaufmann) pp 695-702
- Gullapalli V, Franklin J A, Benbrahim H 1994 Acquiring robot skills via reinforcement learning. *IEEE Contr. Syst. Mag.* : 13-24
- Hertz J A, Krogh A S, Palmer R G 1991 *Introduction to the theory of neural computation* (Reading, MA: Addison-Wesley)
- Jaakkola T, Jordan M I, Singh S P 1994 Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural information processing systems 6* (eds) J D Cowan, G Tesauro, J Alsppector (San Francisco, CA: Morgan Kaufmann) pp. 703-710
- Jacobs R A, Jordan M I, Nowlan S J, Hinton G E 1991 Adaptive mixtures of local experts. *Neural Comput.* 3: 79-87

- Jordan M I, Jacobs R A 1990 Learning to control an unstable system with forward modeling. In *Advances in neural information processing systems 2* (ed.) D S Touretzky (San Mateo, CA: Morgan Kaufmann)
- Jordan M I, Rumelhart D E 1990 Forward models: Supervised learning with a distal teacher. Center for Cognitive Science, Occasional Paper # 40, Massachusetts Institute of Technology, Cambridge, MA
- Kaelbling L P 1990 *Learning in embedded systems*. (Technical Report, TR-90-04) Ph D thesis, Department of Computer Science, Stanford University, Stanford, CA
- Kaelbling L P 1991 *Learning in Embedded Systems* (Cambridge, MA: MIT Press)
- Klopf A H 1972 Brain function and adaptive systems - a heterostatic theory. Technical report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA
- Klopf A H 1982 *The hedonistic neuron: A theory of memory, learning and intelligence*. (Washington, D C: Hemisphere)
- Klopf A H 1988 A neuronal model of classical conditioning. *Psychobiology* 16: 85-125
- Korf R E 1990 Real-time heuristic search. *Artif. Intell.* 42: 189-211
- Kumar P R 1985 A survey of some results in stochastic adaptive control. *SIAM J. Contr. Optim.* 23: 329-380
- Lin C S, Kim H 1991 CMAC-based adaptive critic self-learning control. *IEEE Trans. Neural Networks* 2: 530-533
- Lin L J 1991a Programming robots using reinforcement learning and teaching. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 781-786, MIT Press, Cambridge, MA
- Lin L J 1991b Self-improvement based on reinforcement learning, planning and teaching. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 323-327
- Lin L J 1991c Self-improving reactive agents: Case studies of reinforcement learning frameworks. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behaviour*, (Cambridge, MA: MIT Press) pp 297-305
- Lin L J 1992 Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learning* 8: 293-321
- Lin L J 1993 Hierarchical learning of robot skills by reinforcement. In *Proceedings of the 1993 International Conference on Neural Networks* pp 181-186
- Linden A 1993 On discontinuous Q-functions in reinforcement learning. Available via anonymous ftp from archive.cis.ohio-state.edu in directory /pub/neuroprose

- Maes P, Brooks R 1990 Learning to coordinate behaviour. In *Proceedings of the Eighth National Conference on Artificial Intelligence* (San Mateo, CA: Morgan Kaufmann) pp 796-802
- Magriel P 1976 *Backgammon* (New York: Times Books)
- Mahadevan S, Connell J 1991 Scaling reinforcement learning to robotics by exploiting the subsumption architecture. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 328-332
- Mazzoni P, Andersen R A, Jordan M I 1990  $A_R-P$  learning applied to a network model of cortical area 7a. In *Proceedings of the 1990 International Joint Conference on Neural Networks 2*: 373-379
- Michie D, Chambers R A 1968 BOXES: An experiment in adaptive control. *Machine intelligence 2* (eds) E Dale, D Michie (New York: Oliver and Boyd) pp 137-152
- Millan J D R, Torras C 1992 A reinforcement connectionist approach to robot path finding in non maze-like environments. *Mach. Learning* 8: 363-395
- Minsky M L 1954 *Theory of neural-analog reinforcement systems and application to the brain-model problem*. Ph D thesis, Princeton University, Princeton, NJ
- Minsky M L 1961 Steps towards artificial intelligence. In *Proceedings of the Institute of Radio Engineers* 49: 8-30 (Reprinted 1963 in *Computers and thought* (eds) E A Feigenbaum, J Feldman (New York: McGraw-Hill) pp 406-450
- Moore A W 1990 *Efficient memory-based learning for robot control*. Ph D thesis, University of Cambridge, Cambridge, UK
- Moore A W 1991 Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 328-332
- Moore A W, Atkeson C G 1993 Memory-based reinforcement learning: Efficient computation with prioritized sweeping. In *Advances in neural information processing systems 5* (eds) S J Hanson, J D Cowan, C L Giles (San Mateo, CA: Morgan Kaufmann) pp 263-270
- Mozer M C, Bacharach J 1990a Discovering the structure of reactive environment by exploration. In *Advances in neural information processing 2* (ed.) D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 439-446
- Mozer M C, Bacharach J 1990b Discovering the structure of reactive environment by exploration. *Neural Computation* 2: 447-457
- Narendra K, Thathachar M A L 1989 *Learning automata: An introduction* (Englewood Cliffs, NJ: Prentice Hall)

- Peng J, Williams R J 1993 Efficient learning and planning within the Dyna framework. In *Proceedings of the 1993 International Joint Conference on Neural Networks*, pp 168-174
- Platt J C 1991 Learning by combining memorization and gradient descent. *Advances in neural information processing systems 3* (eds) R P Lippmann, J E Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 714-720
- Rosen B E, Goodwin J M, Vidal J J 1991 Adaptive range coding. *Advances in neural information processing systems 3* (eds) R P Lippmann, J E Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 486-494
- Ross S 1983 *Introduction to stochastic dynamic programming* (New York: Academic Press)
- Rummery G A, Niranjan M 1994 On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, University of Cambridge, Cambridge, England
- Samuel A L 1959 Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* : 210-229 (Reprinted 1963 in *Computers and thought* (eds) E A Feigenbaum, J Feldman (New York: McGraw-Hill))
- Samuel A L 1967 Some studies in machine learning using the game of checkers, II - Recent progress. *IBM J. Res. Dev.* : 601-617
- Selfridge O, Sutton R S, Barto A G 1985 Training and tracking in robotics. In *Proceedings of the Ninth International Joint Conference of Artificial Intelligence* (ed.) A Joshi (San Mateo, CA: Morgan Kaufmann) pp 670-672
- Shepansky J F, Macy S A 1987 Teaching artificial neural systems to drive: Manual training techniques for autonomous systems. In *Proceedings of the First Annual International Conference on Neural Networks*, San Diego, CA
- Singh S P 1991 Transfer of learning across composition of sequential tasks. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 348-352
- Singh S P 1992a Reinforcement learning with a hierarchy of abstract models. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, CA
- Singh S P 1992b On the efficient learning of multiple sequential tasks. In *Advances in neural information processing systems 4* (eds) J E Moody, S J Hanson, R P Lippmann (San Mateo, CA: Morgan Kaufmann) pp 251-258
- Singh S P 1992c Scaling Reinforcement learning algorithms by learning variable temporal resolution models. In *Proceedings of the Ninth International Machine Learning Conference*
- Singh S P 1992d Transfer of learning by composing solutions of elemental sequential tasks. *Mach. Learning* 8: 323-339



- Singh S P, Barto A G, Grunen R, Connelly C 1994 Robust reinforcement learning in motion planning. In *Advances in neural information processing systems 6* (eds) J D Cowan, G Tesauro, J Alspecter (San Fransisco, CA: Morgan Kaufmann) pp 655-662
- Singh S P, Yee R C 1993 An upper bound on the loss from approximate optimal-value functions. Technical Report, University of Massachusetts, Amherst, MA
- Sutton R S 1984 *Temporal credit assignment in reinforcement learning*. Ph D thesis, Univerity of Massachusetts, Amherst, MA
- Sutton R S 1988 Learning to predict by the method of temporal differences. *Mach. Learning* 3: 9-44
- Sutton R S 1990 Integrated architecture for learning, planning, and reacting based on approximating dyanmic programming. In *Proceedings of the Seventh International Conference on Machine Learning* (San Mateo, CA: Morgan Kaufmann) pp 216-224
- Sutton R S 1991a Planning by incremental dynamic programming. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 353-357
- Sutton R S 1991b Integrated modeling and control based on reinforcement learning and dynamic programming. In *Advances in neural information processing systems 3* (eds) R P Lippmann, J E Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 471-478
- Sutton R S, Barto A G 1981 Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol. Rev.* 88: 135-170
- Sutton R S, Barto A G 1987 A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Erlbaum, Hillsdale, NJ
- Sutton R S, Barto A G 1990 Time-derivative models of Pavlovian reinforcement. *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (eds) M Gabriel, J Moore (Cambridge, MA: MIT Press) pp 497-537
- Sutton R S, Singh S P 1994 On step-size and bias in TD-learning. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems* Yale University, pp 91-96
- Sutton R S, Barto A G, Williams R J 1991 Reinforcement learning is direct adaptive optimal control. In *Proceedings of th American Control Conference* Boston, MA, pp 2143-2146
- Tan M 1991 Larning a cost-sensitive internal representation for reinforcement learning. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 358-362
- Tesauro G J 1992 Practical issues in temporal difference learning. *Mach. Learning* 8: 257-278

- Tham C K, Prager R W 1994 A modular Q-learning architecture for manipulator task decomposition. In *Machine Learning: Proceedings of the Eleventh International Conference* (eds) W W Cohen, H Hirsh (Princeton, NJ: Morgan Kaufmann) (Available via gopher from Dept. of Eng., University of Cambridge, Cambridge, England)
- Thrun S B 1986 Efficient exploration in reinforcement learning. Technical report CMU-CS-92-102, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
- Thrun S B 1993 Exploration and model building in mobile robot domains. In *Proceedings of the 1993 International Conference on Neural Networks*
- Thrun S B, Muller K 1992 Active exploration in dynamic environments. In *Advances in neural information processing systems 4* (eds) J E Moody, S J Hanson, R P Lippmann (San Mateo, CA: Morgan Kaufmann)
- Thrun S B, Schwartz A 1993 Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School* (Hillsdale, NJ: Lawrence Erlbaum)
- Tsitsiklis J N 1993 Asynchronous stochastic approximation and Q-learning. Technical Report, LIDS-P-2172, Laboratory for Information and Decision Systems, MIT, Cambridge, MA
- Utgoff P E, Clouse J A 1991 Two kinds of training information for evaluation function learning. In *Proceedings of the Ninth Annual Conference on Artificial Intelligence* (San Mateo, CA: Morgan Kaufmann) pp 596-600
- Watkins 1989 *Learning from delayed rewards*. Ph D thesis, Cambridge University, Cambridge, England
- Watkins C J C H, Dayan P 1992 Technical note: Q-learning. *Mach. Learning* 8: 279-292
- Werbos P J 1987 Building and understanding adaptive systems: a statistical/numerical approach to factory automation and brain research. *IEEE Trans. Syst, Man Cybern.*
- Werbos P J 1988 Generalization of back propagation with application to recurrent gas market model. *Neural Networks* 1: 339-356
- Werbos P J 1989 Neural network for control and system identification. In *Proceedings of the 28th Conference on Decision and Control* Tampa, FL, pp 260-265
- Werbos P J 1990a Consistency of HDP applied to simple reinforcement learning problems. *Neural Networks* 3: 179-189
- Werbos P J 1990b A menu of designs for reinforcement learning over time, In *Neural networks for control* (eds) W T Miller, R S Sutton, P J Werbos (Cambridge, MA: MIT Press) pp 67-95

- Werbos P J 1992 Approximate dynamic programming for real-time control and neural modeling. In *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches* (eds) D A White, D A Sofge (New York: Van Nostrand Reinhold) pp 493-525
- Whitehead S D 1991a A complexity analysis of cooperative mechanisms in reinforcement learning. In *Proceedings of the Ninth Conference on Artificial Intelligence*, (Cambridge, MA: MIT Press) pp 607-613
- Whitehead S D 1991b Complexity and cooperation in Q-learning. In *Machine Learning: Proceedings of the Eighth International Workshop* (eds) L A Birnbaum, G C Collins (San Mateo, CA: Morgan Kaufmann) pp 363-367
- Whitehead S D, Ballard D H 1990 Active perception and reinforcement learning. *Neural Comput.* 2: 409-419
- Williams R J 1986 Reinforcement learning in connectionist networks: a mathematical analysis. Technical report ICS 8605, Institute for Cognitive Science, University of California at San Diego, La Jolla, CA
- Williams R J 1987 Reinforcement-learning connectionist systems. Technical report NU-CCS-87-3, College of Computer Science, Northeastern University, Boston, MA
- Williams R J, Baird L C III 1990 A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming. In *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems* New Haven, CT, pp 96-101



# Dynamic scheduling in manufacturing systems using Brownian approximations

K RAVIKUMAR and Y NARAHARI

Department of Computer Science and Automation  
Indian Institute of Science, Bangalore

**Abstract.** Recently, Brownian networks have emerged as an effective stochastic model to approximate multiclass queueing networks with dynamic scheduling capability, under conditions of balanced heavy loading. This paper is a tutorial introduction to dynamic scheduling in manufacturing systems using Brownian networks.. The article starts with motivational examples. It then provides a review of relevant weak convergence concepts, followed by a description of the limiting behaviour of queueing systems under heavy traffic. The Brownian approximation procedure is discussed in detail and generic case studies are provided to illustrate the procedure and demonstrate its effectiveness. This paper places emphasis only on the results and aspires to provide the reader with an up-to-date understanding of dynamic scheduling based on Brownian approximations.

**Keywords.** Brownian networks; dynamic scheduling; manufacturing systems; multiclass queueing networks; heavy traffic approximations; weak convergence; functional central limit theorem.

## 1. Introduction

Scheduling as a research area is motivated by important resource allocation questions that arise in manufacturing systems, computer systems, computer communication networks, and in general, in all situations where scarce resources have to be allocated to activities over time to appropriate servers (processors, machines, communication channels, material handling devices, etc.) so as to optimize a performance criterion, while satisfying a set of given constraints. Scheduling problems can be classified as *static* scheduling problems when the jobs to be scheduled comprise a fixed set and *dynamic* when jobs can arrive into the facility in an ongoing and usually, in a random fashion. Another usual way of classifying scheduling problems is to consider them as *deterministic* or *stochastic*. In *deterministic* scheduling, job characteristics such as processing times, due dates, and release dates are known with certainty to the scheduler before the actual processing occurs. In *stochastic* scheduling, the scheduler cannot observe the processing times in advance, but only has knowledge

of a probability distribution for the various processing times. In this paper, the emphasis is on *dynamic* and *stochastic scheduling* of multi-class queueing network models of discrete event systems, using a class of *heavy traffic* approximations, called **Brownian approximations**. Also all our motivating scheduling problems come from the area of manufacturing systems, though the methodology that we discuss is applicable, in general, to any discrete activity scheduling problem with dynamic and stochastic characteristics.

### 1.1 Deterministic scheduling

Much of the research in the area of scheduling has focussed on deterministic scheduling problems. Most of the scheduling problems in this area have been shown to be NP-hard and researchers have explored several different approaches to confront NP-hardness.

- Determine a *lower bound* on the cost of the schedule and then use a *branch and bound* method to determine the optimal solution (Bagchi & Ahmadi 1987; Beloudah *et al* 1988). However, this technique needs exponential amount of computation time in the worst case.
- Use *dynamic programming* (Abdul-Razaq & Potts 1988; Baker & Ahmadi 1978). This technique works very well for many scheduling problems, but like branch and bound technique, needs exponential amount of computation time in the worst case.
- Obtain *sub-optimal* solutions in polynomial time (Hochbaum & Shmoys 1988). Such approximation algorithms are, however, applicable only in specific problem instances and do not yield general methods.
- Use simple *heuristics* (Gere 1987) such as EDD (Earliest Due Date), SPT (Shortest Processing Time), etc. Heuristics are very efficient and have the ability to react to dynamic changes and have widespread applicability. In general, however, heuristics do not offer the guarantee that the solution is within an acceptable margin of error when compared with the optimal solution.
- *Lagrangian relaxation* based methods (Fisher 1973, 1981) which yield efficient near-optimal solutions with measurable performance as well as important job interaction information to accommodate dynamic changes and to handle new jobs.
- More recently, randomized local search algorithms such as *simulated annealing* (Van Laarhoven *et al* 1992) and *genetic algorithms* (Goldberg 1986) have been applied to deterministic scheduling problems. Another paradigm that has also been used in this context is *neural networks* (Levy & Adams 1987).

### 1.2 Stochastic scheduling

In the area of stochastic scheduling, the results are scattered and technically complicated (Lawler *et al* 1990); they rely on semi-Markovian decision theory and stochastic dynamic optimization. Important results in this dynamic optimization are sur-

veyed by Lawler *et al* (1990), Weiss (1982), Pinedo and coworkers (Pinedo & Weiss 1980, 1987; Pinedo 1981-1983, Pinedo & Scragg 1982) and Forst (1984).

Single class and multiclass queueing networks constitute an important class of stochastic models of discrete event systems (see Walrand 1988). The optimal scheduling of such networks has been attempted by several researchers, but only with limited success. Some of the notable efforts in this area include:

- Priority sequencing in single station queueing systems (Klimov 1974)
- Optimal dynamic scheduling in Jackson networks (Ross & Yao 1989)
- Optimal scheduling control in a flexible machine (Yao & Shantikumar 1990)
- Optimal control of interacting service stations (Hajek 1974)
- Optimal control of service rates in networks of queues (Weber & Stidham 1987)
- Optimal control of admission to a queueing system (Stidham 1985)

However, according to Harrison & Wein (1989), a satisfactory theory for sequencing and scheduling in a queueing network setting has yet to be formulated. Discrete event simulation continues to be the primary tool of analysis and the best hope for further progress appears to be in the analysis of cruder and more tractable models.

Recently, *Brownian networks* (Harrison 1988) have emerged as an effective stochastic model to approximate multiclass queueing networks with dynamic scheduling capability, under conditions of *balanced heavy loading* (see §3). A Brownian network is a crude model but highly tractable and successful in the context of dynamic and stochastic scheduling of queueing networks. This paper attempts to survey the important results in this area. In particular, we present:

- foundational aspects of Brownian networks as applied to the modeling of multiclass queueing networks,
- methodological details of how sequencing and scheduling problems can be approached via the Brownian approximation,
- several illustrative case studies to gain insight into specific methodological details.

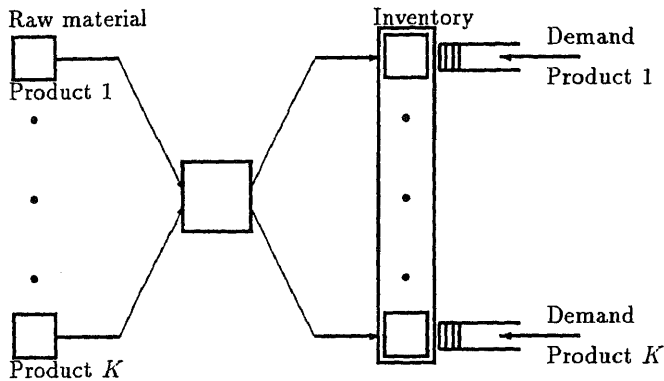
Since this paper is intended as a tutorial review, we have used extensively the results from many important papers in this area. These papers include: Harrison (1988), Harrison & Wein (1990), and Wein (1990, 1992). Wherever highly relevant, we shall again explicitly provide a reference to these papers.

### 1.3 Motivational examples

In this section we describe some scheduling problems that occur in dynamic and stochastic environments through some simple and illustrative manufacturing system examples.

### 1.3.1 A multiclass make-to-stock queue

A make-to-stock production facility produces a wide variety of products according to a forecast of customer demands and completed jobs enter a finished good inventory, which in turn services the actual customer demand. Figure 1 shows a single machine make-to-stock system with  $K$  classes of products.



**Figure 1.** A single machine make-to-stock system.

We shall consider a system with  $K = 5$ . The class designations summarize all relevant information with regard to processing times and demand patterns of respective products as given below:

- *Class 1:* Product's processing time is low but demands arrive frequently for it.
- *Class 2:* Products have long processing times but demands arrive occasionally for them.
- *Class 3:* These are high priority products with medium processing times and nominal demands. Waiting times of customers arriving for these products should be low.
- *Class 4:* These are products with medium processing times but very high demands.
- *Class 5:* For these products processing times are medium and demands are occasional.

An arriving class  $k$  customer takes a product of the same class, if available; otherwise, backorders for one. Linear costs per unit time are incurred for holding inventory and for backordering.

In a realistic scenario, actual processing times are not known with certainty and one can only have a knowledge of probability distributions of various processing times. It is common that every such manufacturing system experiences some amount



of variability in estimated processing times. Also, sometimes these variations turn out to be unpredictable. For example, variations due to rework, machine failure *etc.* fall under this category. These variations will have impact on costs incurred by the system. For instance, long processing times may reduce inventory holding costs but at the same time they incur high backordering costs. Also, variations in interarrival times of demands produce similar effects. Hence, given the stochastic nature of the problem, deterministic scheduling is less realistic than dynamic scheduling. Further, scheduling policies which perform well in the deterministic setting may not perform well in stochastic setting.

A typical dynamic scheduling decision for the foregoing problem consists of choosing among the following options at each point in time:

- either work on a class  $k$  job,  $k = 1, \dots, 5$
- or allow the machine to idle.

Using Brownian analysis methodology, Wein (1992) derived a dynamic scheduling policy for a make-to-stock system under general service distributions and renewal demand patterns. The decision as to whether a machine is to be kept busy or idle at any time point is dictated by the weighted inventory level process (which is a weighted sum of inventory levels of each class, the weights being mean processing times.) at that time. The priority decision derived is reminiscent of the well known  $c_k \mu_k$  - rule, which awards priority to the class with the largest value of the index  $c_k \mu_k$  where  $c_k$  is the holding cost and  $\mu_k$  is the service rate. This policy is discussed in detail in §4.1. Simulation results showing comparison of this policy with other conventional policies are also presented.

### 1.3.2 A re-entrant line

Re-entrant lines are queueing network models for wafer fabrication in a semi-conductor manufacturing system. Wafer fabrication involves a large and complex sequence of processing steps. A characteristic feature of wafer fabrication is *re-entrancy*, that is each wafer visits the same machine centre multiple number of times. Figure 2 depicts a three-station re-entrant line with a single job type.

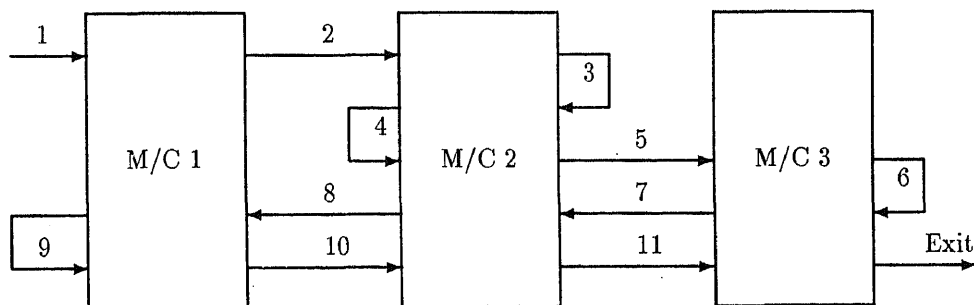


Figure 2. A three station re-entrant line.

Each job has its own deterministic route through the network. We can define a different customer class for each operation of each job. For example, in figure 2, a job has to undergo 11 stages of operation and hence, has eleven classes associated with it. Each job class has its own processing time distribution and different classes contend for service at the same machine center. If the population size of jobs circulating in the system is held constant, *i.e.*, a new job is admitted whenever a job leaves the network, then the above system can be modeled as a three station multiclass closed queueing network. A scheduling problem of relevance in this context is to choose a policy which, at each point in time indicates which class to be serviced at each station. Some of the conventional policies employed in scheduling a re-entrant line are FCFS, FBFS (First Buffer First Served), LBFS (Last Buffer First Served), SEPT (Shortest Expected Processing Time) *etc.*

In re-entrant lines, processing times of various operations are susceptible to unpredictable variability mainly because of complexity and precision requirements involved in performing the operations. Furthermore, due to multiple visits of jobs, many machine centres will be heavily loaded and thus become bottlenecks. In the example shown in figure 2, stations 1 and two are bottleneck stations. These bottlenecks are precisely where large queues form, where most of the waiting time is incurred and where scheduling will have biggest impact. Hence, under such scenario, utilizations of the bottleneck machines should be as high as possible to reduce cycle times and this can be affected through scheduling decisions which take into account the state of the system at any point in time.

A similar scheduling problem for two station closed queueing network is considered in Harrison & Wein (1990) and the scheduling decision considered there is referred to as *workload balancing rule*. It is a static priority sequencing policy, which assigns priority at any station according to an index rule which aims at minimizing workload imbalance between the two stations and there by enhancing the utilization levels of the machines.

This policy works well when the network has more than one bottleneck stations and not too many non-bottleneck stations. If the network has only one bottleneck it is difficult to affect utilizations of bottleneck machines because there are no other bottleneck stations to feed it. Similarly, presence of too many non-bottleneck stations prevent bottleneck stations to feed one another in an effective manner. Details of this policy are given in §4.3. A simulation study is conducted using the workload balancing policy, on the above re-entrant line examples. Section 4.3 also contains results of these experiments.

In the example of figure 2, one can easily see that by admitting a new job into the network, the number of customers of each class goes up by one. Thus, if an open loop release policy which pushes jobs into the system without observing the state of the system is followed then the WIP (Work-In-Process) inventory levels shoot up drastically. From Little's law, it is known that for a given mean arrival rate, mean cycle times are directly proportional to mean WIP. However, the relationship between mean throughput rate and mean WIP is highly non-linear and dependent on the scheduling policy. Using an effective job release policy and a priority sequencing policy combination one can achieve high throughput rates while maintaining low levels of WIP.

An interesting job release policy, known as workload regulating policy, is consid-

ered in Wein (1990b). This policy injects a job into the system whenever the amount of work in the system for the bottleneck stations satisfies certain conditions. The priority sequencing policy uses dynamic reduced costs from a linear program. These policies along with some results of simulation experiment performed on the above reentrant line are presented in §4.4.

The remainder of this paper is organized as follows. In §2.1 we describe weak convergence concepts of relevance and discuss in §2.2, how the heavy traffic limit theorems are proved invoking these concepts. The description of the Brownian network, followed by the approximation procedure, is given in §3.1. Section 3.2 provides workload formulation for the Brownian network of §3.1, which describes the system dynamics of the queueing network in terms of workloads at service centres. Modifications needed to adopt the approximation procedure to the case of closed queueing networks are presented in §3.3. Section 4 illustrates the procedure in the context of various manufacturing systems of practical importance. Section 4.1 deals with a scheduling problem in a single machine make-to-stock queue. Section 4.2 discusses the case of a two-station closed queueing network, with an objective to maximize the throughput. Section 4.3 gives an interesting scheduling problem in a two-station network with controllable inputs. Here we mention that no attempt is made to provide rigorous proofs for the theorems presented and interested readers are referred to appropriate contributions for a detailed study of the problem concerned.

## 2. Foundations

### 2.1 Weak convergence concepts

In this section we describe some relevant notions of weak convergence which will be used in subsequent portions of this paper.

Let  $\{X(t), t \in T\}$  be a stochastic process on a probability space  $(\Omega, \mathcal{E}, \mathcal{P})$ . Suppose that the index set  $T$  is an interval of the real line  $R$ . For a fixed  $\omega \in \Omega$ , the function  $X(\omega, \cdot)$  of  $t$  gives a sample path of the process. If all such sample paths lie in some fixed collection  $\mathcal{X}$  of real valued functions on  $T$ , then the process  $X$  can be thought of as a map from  $\Omega$  into  $\mathcal{X}$ , a random element of  $\mathcal{X}$ . For example, if a process indexed by  $[0, 1]$  has continuous sample paths it will be a random element of the space  $C[0, 1]$  of all real-valued continuous functions on  $[0, 1]$ . However, the notion of random element needs to be formalized adding measurability requirement as follows:

**DEFINITION 2.1.1** An  $\mathcal{E}/\mathcal{A}$ -measurable map  $X$  from a probability space  $(\Omega, \mathcal{E}, \mathcal{P})$  into a set  $\mathcal{X}$  with a  $\sigma$ -field  $\mathcal{A}$  is called **random element** of  $\mathcal{X}$ .

If  $\mathcal{X}$  is a metric space, the set of all bounded, continuous  $\mathcal{A}/\mathcal{B}(R)$  measurable, real-valued functions on  $\mathcal{X}$  is denoted by  $\mathcal{C}(\mathcal{X}, \mathcal{A})$ . Note that if  $\mathcal{A}$  is the Borel field generated by closed sets of  $\mathcal{X}$ , then every continuous function on  $\mathcal{X}$  is measurable. A sequence  $\{X_n\}$  of random elements of  $\mathcal{X}$  converges in distribution to a random element  $X$ , written as  $X_n \Rightarrow X$ , if

$$\int f(X_n) d\mathcal{P} \longrightarrow \int f(X) d\mathcal{P} \text{ for each } f \in \mathcal{C}(\mathcal{X}, \mathcal{A}) \quad (1)$$

A sequence  $\{P_n\}$  of probability measures on  $\mathcal{A}$  converge weakly to  $P$ , written as  $P_n \Rightarrow P$  if

$$\int f dP_n \longrightarrow \int f dP \text{ for each } f \in \mathcal{C}(\mathcal{X}, \mathcal{A}) \quad (2)$$

As every random element  $X$  of  $\mathcal{X}$  induces a probability measure  $P$  on  $(\mathcal{X}, \mathcal{A})$  defined by

$$P(A) = \mathcal{P}(X^{-1}(A)), \text{ for all } A \in \mathcal{A} \quad (3)$$

convergence in distribution of a sequence of random elements is synonymous to weak-convergence of the corresponding sequence of induced probability measures. However, note that in the latter case  $X_n$  and  $X$  need not be defined on a same probability space but must induce probability measures  $P_n$  and  $P$  on the same metric space  $(\mathcal{X}, \mathcal{A})$ .

Now on, unless otherwise stated, assume that  $\mathcal{X}$  is a separable metric space with metric  $\rho$  and the Borel  $\sigma$ -field  $\mathcal{A}$ . If  $X$  and  $Y$  are defined on a common domain, then  $\rho(X, Y)$  is a random variable (see Billingsley 1968). Thus the following definition makes sense.

**DEFINITION 2.1.2** A sequence of random elements  $\{X_n, n \geq 1\}$  converges in probability to  $X$ , written as  $X_n \xrightarrow{\mathcal{P}} X$ , if  $X_n$  and  $X$  are defined on a common probability space  $(\Omega, \mathcal{E}, \mathcal{P})$  and

$$\rho(X_n, X) \xrightarrow{\mathcal{P}} 0.$$

Here  $\xrightarrow{\mathcal{P}}$  denotes convergence in probability of random variables.

Now we state a useful theorem whose application is found frequently in weak convergence results for queueing theory.

**Theorem 2.1.1** Assume that  $\{X_n\}$  and  $\{Y_n\}$  are sequences of random elements of  $\mathcal{X}$  and are defined on a common probability space  $(\Omega, \mathcal{E}, \mathcal{P})$ . If  $X_n \Rightarrow X$  and  $\rho(X_n, Y_n) \xrightarrow{\mathcal{P}} 0$ , then  $Y_n \Rightarrow X$ .

Stochastic processes of interest in queueing theory such as queue length process can often be represented as functions of more basic stochastic processes such as random walks and renewal processes. Consequently limit theorems for stochastic processes in queueing theory are often obtained from existing limit theorems for these basic processes by showing that the connecting functions preserve convergence. The functions that appear in such proofs are composition, addition, multiplication, supremum, etc. Hence, a natural question that arises in such contexts is: If  $X_n \Rightarrow X$  and  $f$  is a measurable mapping from  $(\mathcal{X}, \mathcal{A})$  to another separable metric space  $(\mathcal{X}', \mathcal{A}')$ , does it follow that  $f(X_n) \Rightarrow f(X)$ ? Observe that the result is trivially true if  $f$  is continuous. Interestingly this holds even under slightly weaker assumption as shown by the following theorem:

**Theorem 2.1.2** (Continuous mapping theorem) If  $X_n \Rightarrow X$  and  $f$  is continuous almost surely with respect to the distribution of  $X$ , then  $f(X_n) \Rightarrow f(X)$ .

The above theorem can be further generalized as given below.

**Theorem 2.1.3** Let  $f_n, n \geq 1$  and  $f$  be Borel measurable functions mapping the separable metric space  $(\mathcal{X}, \mathcal{A})$  into another separable metric space  $(\mathcal{X}', \mathcal{A}')$ . If  $X_n \Rightarrow X$  and  $f_n(x_n) \rightarrow f(x)$  for all  $x \in A$  and  $\{x_n \rightarrow x\}$ , then  $f_n(X_n) \Rightarrow f(X)$ .

Most of the weak convergence results in queueing theory rest on the continuous mapping theorem. For an elegant proof of this theorem see Pollard (1984) or Whitt (1980). In queueing theory, the metric spaces of particular interest are  $C[0,1]$ , the space of all real-valued continuous functions on  $[0,1]$  and the space  $D[0,1]$  of all real-valued functions that are right continuous at each point of  $[0,1]$  with left limits existing at each point of  $(0,1]$ . The functions of  $D[0,1]$  are called *cadlag* functions.

Thus, the space  $D[0,1]$  contains the sample paths of all queue-related processes. Obviously,  $C[0,1] \subset D[0,1]$ .

The metric on  $C[0,1]$  is the uniform metric defined by

$$\rho(x, y) = \sup_{0 \leq t \leq 1} |x(t) - y(t)| \text{ for all } x, y \in C[0,1].$$

Under this metric  $\rho$ ,  $C[0,1]$  is complete and separable. But under the same uniform metric  $D[0,1]$  is complete but not separable and hence the uniform metric poses some minor measurability difficulties. For instance, under this metric the Borel  $\sigma$ -field turns out to be too large and many interesting stochastic processes fail to be random elements of  $D[0,1]$ . However, if we consider a strictly smaller  $\sigma$ -field  $\mathcal{B}$  generated by closed balls, an interesting weak convergence theory results.  $\mathcal{B}$  also coincides with the  $\sigma$ -field generated by co-ordinate projection maps. All interesting functionals on  $D[0,1]$  are  $\mathcal{B}$  measurable. The lack of a countable dense subset of functions in  $D[0,1]$  is surmounted when the limit distributions concentrate on a separable subset of  $D[0,1]$  such as  $C[0,1]$ . For an interesting theory under the uniform metric, see Pollard (1984).

If in the space  $D[0,1]$ , we define  $f_u(t) = I\{t \geq u\}$ ;  $t \in [0,1]$  and  $u \in [0,1]$ , the collection  $\{f_u(\cdot)\}$  is uncountable and under uniform metric two distinct functions are at unit distance. Clearly we want  $f_u \rightarrow f_v$  whenever  $u \rightarrow v$ . However, this is quite impossible under any topology that leads to a convergence concept which implies pointwise convergence. Skorohod surmounted this difficulty by defining a metric, which leads to pointwise convergence after a suitable rescaling of the time axis that becomes asymptotically negligible. It is difficult to define this metric. It suffices for our purpose to know that such a metric exists and under this metric  $D[0,1]$  is separable. Unfortunately, under this metric the space is not complete. Billingsley (1968) defines an equivalent metric under which  $D[0,1]$  is both separable and complete. Here onwards, we concentrate on  $D[0,1]$  (hereafter to be denoted as  $D$ ) equipped with the metric under which it is both separable and complete.

Most of the diffusion approximations and heavy traffic limit theorems rely on the so called *Functional central limit theorem*. The functional central limit theorem is an extension of classical Lindberg-Levy's central limit theorem (see Breiman 1968) for a sequence of random variables, to the function space  $D$ . It is directed at showing that a normalized sequence of random functions converges to a diffusion process. The advantage of functional limit theorems lies in the fact that weak convergence results can be immediately obtained for various functionals of the processes. The standard method to prove weak convergence of a normalized sequence of processes is to first show the convergence of finite dimensional distributions. However, this is not sufficient. A certain *tightness* property of the induced measures needs to be demonstrated and this part poses some technical difficulties.

To get a feel for weak convergence in function spaces, in what follows we state various functional limit theorems of relevance in queueing theory.

**Theorem 2.1.4 (Donsker's theorem)** Let  $\{\xi_i, i \geq 1\}$  be a sequence of i.i.d. random variables with mean 0 and variance  $\sigma^2 < \infty$ , defined on  $(\Omega, \mathcal{E}, \mathcal{P})$ . Let

$$S_k = \xi_1 + \cdots + \xi_k \text{ for all } k \geq 1 \text{ and } S_0 = 0$$

From the  $\{S_n\}$ , form random elements  $\{X_n\}$  of  $D$  as:

$$X_n(t) = \frac{S_{[nt]}}{\sigma\sqrt{n}} \text{ for all } t \in [0, 1]$$

where  $[x]$  is the greatest integer less than or equal to  $x$ . Let  $W = \{W_t, 0 \leq t \leq 1\}$  be the standard Brownian motion. Then

$$X_n \Rightarrow W \text{ in } D$$

For a proof of this see Billingsley (1968).

A generalization of the Donsker's theorem is due to Prohorov and is given by:

**Theorem 2.1.5 (Prohorov's theorem)** Suppose for each  $n \geq 1$  there exists a sequence of i.i.d. random variables  $\{\eta_i^n, i \geq 1\}$  with mean 0 and variance  $\sigma^2$ . Define partial sums

$$S_k^n \equiv \eta_1^n + \dots + \eta_k^n \text{ and } S_0^n \equiv 0.$$

Assume that

$$\sigma_n^2 \rightarrow \sigma^2 \text{ as } n \rightarrow \infty, 0 < \sigma^2 < \infty$$

and

$$\sup_{n \geq 1} E\{|\eta_1^n|^{2+\epsilon}\} < \infty \text{ for some } \epsilon > 0$$

Let

$$X_n(t) \equiv \frac{S_{[nt]}^n}{\sigma\sqrt{n}}.$$

Then,

$$X_n \Rightarrow W \text{ in } D.$$

In the above theorems the partial sums  $S_n$  are defined for fixed indices  $n$ . Sometimes we encounter cases where the index is random and such a phenomenon is common in renewal processes. Suppose  $\nu_n$  is a random integer such that  $\nu_n$  is large with high probability. Define random elements  $X_n$  of  $D$  as given in Donsker's theorem. Further, define another random element  $Y_n$  of  $D$  by,

$$Y_n(t) = \frac{S_{\nu_{[nt]}(\omega)}(\omega)}{\sigma\sqrt{n}}$$

Now we seek conditions under which  $\{Y_n\}$  weakly converges to some limit. Observe that  $Y_n(\omega)$  results from  $X_n(\omega)$  by subjecting  $X_n$  to a random time scale. If we define  $\phi_n(t, \omega)$  by

$$\phi_n(t, \omega) = \nu_{[nt]}(\omega)/n$$

then it follows that

$$Y_n(t, \omega) = X_n(\phi_n(t, \omega), \omega).$$

Thus  $Y_n$  is  $X_n$  with the time scale subjected to a change represented by random function  $\phi_n$ . Such cases can be dealt with using the following theorem, known as the *Random time change theorem*.

Let  $D_0$  denote the set of elements  $\phi$  of  $D$  that are non-decreasing and satisfy  $0 \leq \phi(t) \leq 1$  for all  $t \in [0, 1]$ . For  $X \in D$ ,  $\phi \in D_0$ , let  $(X \circ \phi)(t) = X(\phi(t))$ . Suppose that in addition we have random elements  $X_n$  and  $\phi_n$  of  $D$  and  $D_0$  respectively, where  $X_n$  and  $\phi_n$  have the same domain (which can vary with  $n$ ). Note that  $X \circ \phi$  and  $X_n \circ \phi_n$  for each  $n$  lie in  $D$ . If  $D_0$  is topologized by relativizing the Skorohod topology of  $D$ , then it is easy to see that  $(X, \phi)$  and  $(X_n, \phi_n)$  are random elements of  $D \times D_0$  with product topology. The following result is given by Billingsley (1968).

**Theorem 2.1.6 (Random time change theorem)** *If  $(X_n, \phi_n) \Rightarrow (X, \phi)$  and  $P(X \in C) = P(\phi \in C) = 1$ , then*

$$X_n \circ \phi_n \Rightarrow X \circ \phi$$

where  $C \equiv C[0, 1]$ .

The proof of the above theorem is based on continuous mapping theorem and also on the fact that Skorohod topology relativized to  $C$  coincides with the topology generated by uniform metric on  $C$ . The theorem is useful in deriving functional central limit theorem for renewal processes.

**Theorem 2.1.7** *Let  $\eta_1, \eta_2, \dots$  be an i.i.d. sequence of random variables with mean  $\mu$  and variance  $\sigma^2 < \infty$ . Define,*

$$\nu_t = \max\{k : \sum_{i=1}^k \eta_i \leq t\}, \text{ with } \nu_t = 0 \text{ if } \eta_1 \geq t$$

*Thus  $\nu_t$  gives number of renewals upto time  $t$ . Define*

$$Z_n(t, \omega) = \frac{\nu_{nt}(\omega) - nt/\mu}{\sigma\mu^{-3/2}\sqrt{n}}.$$

*Then,  $Z_n \Rightarrow W$  in  $D$ .*

For a proof of this refer Billingsley (1968).

These notions of weak convergence are used in proving heavy traffic limit theorems for queue related processes as we shall see in §2.3.

## 2.2 Brownian motion

As discussed in the previous section the limit process in the functional central limit theorem is the standard Brownian motion. In this section we define the standard Brownian motion and the reflected Brownian motion which is a functional of Brownian motion. Many of the interesting queue related processes converge to the latter.

**DEFINITION 2.2.1** A standard Brownian motion or Wiener process is a stochastic process  $\{X(t), 0 \leq t \leq 1\}$  on  $(\Omega, \mathcal{E}, \mathcal{P})$  having continuous sample paths, and stationary independent increments such that for any fixed  $t \in [0, 1]$ ,  $X(t)$  is normally distributed with mean 0 and variance  $t$ .

Thus, a standard Brownian motion starts at level zero almost surely.

**DEFINITION 2.2.2** A process  $\{Y(t), 0 \leq t \leq 1\}$  is called a  $(\mu, \sigma)$  Brownian motion if it has the form:

$$Y(t) = Y(0) + \mu t + \sigma X(t). \quad (4)$$

where  $X(t)$  is the standard Brownian motion and  $Y(0)$  is independent of  $X$ .

It follows that  $Y(t+s) - Y(t) \sim N(\mu s, \sigma^2 s)$ .  $\mu$  is called the drift and  $\sigma^2$  the variance of  $Y(t)$ .

The normality requirement in the above definition is superfluous because if  $Y$  is a continuous path process and has independent increments, then  $Y$  is a Brownian motion and normality follows as a consequence of these assumptions. Refer Breiman (1968) and Cox & Miller (1965) for further properties of Brownian motion.

**DEFINITION 2.2.3** Let  $f : D \rightarrow D$  be defined for all  $Y \in D$  as  $f(Y) = Z$  where

$$Z(t) = Y(t) - \inf_{0 \leq s \leq t} \{Y(s)\}, 0 \leq t \leq 1,$$

where  $Y(t)$  is  $(\mu, \sigma)$ Brownian motion with  $Y(0) = 0$ . Then  $\{Z(t), 0 \leq t \leq 1\}$  is called *reflected Brownian motion*, denoted by  $\text{RBM}(\mu, \sigma)$ .

Whitt (1980) proves that  $f$  above is continuous in Skorohod topology.

For processes in  $R^K$ , the limit processes of the functional central limit theorem is a  $K$ -dimensional Brownian motion,  $\{\bar{Y}(t), 0 \leq t \leq 1\}$ , specified by a  $K$ -dimensional drift vector,  $\bar{c}$  and a  $K \times K$  covariance matrix  $A$ , denoted by  $\text{BM}(\bar{c}, A)$ , i.e.,  $\bar{Y}(t)$  is a  $K$ -dimensional vector stochastic process with continuous sample paths in  $R^K$  with  $\bar{Y}(0)=0$  and stationary independent increments.

Similarly, a reflected Brownian motion on the non-negative orthant  $R_+^K$  was defined and characterized by Harrison & Reiman (1981) and is discussed in detail by Harrison & Williams (1987). It behaves like Brownian motion on the interior of its state space  $R_+^K$  and reflects instantaneously in a fixed direction at each boundary hyperplane. The reflection directions are given in  $K \times K$  reflection matrix  $R$ , where the  $k$ -th row of  $R$  gives the reflection direction for the boundary corresponding to  $X_k(t) = 0$ . This process is thus completely specified by  $(\bar{c}, A, R)$  where  $\bar{c}$  and  $A$  correspond to the drift vector and covariance matrix of the underlying Brownian motion.

### 2.3 Heavy traffic limit theorems

A queueing system is stable if the input rate is less than the output rate, i.e., if the traffic intensity,  $\rho$  is strictly less than unity. If  $\rho \geq 1$ , the system is unstable and the queueing processes tend to blow up. For example, in a GI/G/1 queue, if  $\rho \geq 1$ , then for any  $K < \infty$ ,

$$\lim_{n \rightarrow \infty} \mathcal{P}\{W_n \geq K\} = 1,$$

where  $W_n$  is the waiting time of the  $n$ -th customer. Thus if  $\rho \geq 1$  the queue is said to be under heavy traffic. However, even under heavy traffic conditions, properly normalized sequences of queueing processes converge weakly to diffusion processes. Heavy traffic limit theorems formalize this fact. The diffusion approximation procedures stem from these limit theorems.

In this section, we discuss a simple case of GI/G/1 queue under heavy traffic and give an intuitive feel for how the heavy traffic limit theorems are proved invoking the weak convergence concepts discussed in §2.1.

Consider a standard GI/G/1 queue determined by two independent sequences of i.i.d. random variables  $\{u_n, n \geq 1\}$  and  $\{v_n, n \geq 0\}$ . Assume that the 0-th customer arrives at time  $t = 0$  to find a free server. Let  $v_n$  represent the service time of the  $n$ -th customer and  $u_n$  represent the inter-arrival time between the  $(n - 1)$ -st customer



and  $n$ -th customer. We define,

$$\rho = E(v_1)/E(u_1)$$

$$Y_n = v_{n-1} - u_n$$

$$W_{n+1} = [W_n + Y_{n+1}]^+, \text{ for all } n \geq 0 \text{ and } W_0 = 0.$$

$W_n$  gives the waiting time of the  $n$ -th customer. It is well known that if  $\rho < 1$ , then there exists a non-degenerate random variable  $W$  such that :

$$W_n \Rightarrow W \text{ as } n \rightarrow \infty.$$

Under appropriate moment conditions one can show that

$$(\sigma^2 n)^{-1/2} \left[ \sum_{k=1}^n W_k - n E(W) \right] \Rightarrow N(0, 1).$$

In this case, the events  $\{W_k=0\}$  are regenerative points for  $\{W_n, n \geq 1\}$  and  $\{W_k=0\}$  occurs infinitely often, w.p.1. Thus,  $\{W_n, n \geq 1\}$  is a regenerative process and  $\{\sum_{k=1}^n W_k, n \geq 1\}$  is a cumulative process. Thus,  $\{W_n\}$  can be broken up into i.i.d. blocks and consequently, eventhough  $\{W_n\}$  is itself not i.i.d., the theory of sequence of i.i.d. random variables can be applied for a proof of the above convergence.

But, in the case when  $\rho = 1$ , the situation is more delicate and in the context of Markov chains this case corresponds to null recurrence. If  $\rho = 1$ ,  $W_n \leq \mathcal{K}$  for  $\mathcal{K}$  finite, infinitely often w.p.1. But the expected time between epochs when customers arrive to find a free server is infinite. However, observe that,

$$W_n = S_n - \min\{S_k, 0 \leq k \leq n\}, \quad n \geq 0 \quad (5)$$

where  $S_n = \sum_{k=1}^n Y_k$  and  $S_0 = 0$ .

It is apparent that the limit behaviour of  $\{W_n\}$  is closely related to the limit behaviour of  $\{S_n\}$  and not the same, because  $W_n$  is a function of the initial segment  $\{S_k, 0 \leq k \leq n\}$  and not just the single  $S_n$ . This relation between initial segments  $\{W_k, 0 \leq k \leq n\}$  and  $\{S_k, 0 \leq k \leq n\}$  can be established by inducing, for each  $n$ , an appropriate stochastic process in  $D$ . Let:

$$\widetilde{W}_n \equiv \widetilde{W}_n(t) = \frac{W_{[nt]}}{a_n}, \quad 0 \leq t \leq 1$$

$$\widetilde{S}_n \equiv \widetilde{S}_n(t) = \frac{S_{[nt]}}{a_n}, \quad 0 \leq t \leq 1.$$

where  $a_n$  is a normalizing constant such that  $a_n \rightarrow \infty$  as  $n \rightarrow \infty$ .

$\widetilde{W}_n$  and  $\widetilde{S}_n$  are continuous time processes with sample paths in  $D$ . It is apparent from (5) that:

$$\widetilde{W}_n = f(\widetilde{S}_n)$$

where,  $f : D \rightarrow D$  is defined by,

$$f(X)(t) = X(t) - \inf_{0 \leq s \leq t} X(s), \quad 0 \leq t \leq 1.$$

Hence the desired limit theorems follow from the Donsker's theorem and continuous mapping theorem. Thus,  $\tilde{S}_n$  converges weakly to a Brownian motion and hence  $\tilde{W}_n$  converges to a reflected Brownian motion.

For the case when  $\rho > 1$ ,  $\min\{S_k, 0 \leq k \leq n\}$  in equation 5 converges weakly to a non-degenerate random variable. The limiting behaviour of  $W_n$  is obtained from the known results for random walks because from the convergence together theorem it follows that with normalization  $\{W_n\}$  and  $\{S_n\}$  have the same limiting behaviour.

Thus, heavy traffic limit theorems for queueing processes are proved in general by expressing them as functions of some basic processes for which limit theorems exist and then invoking theorems such as the continuous mapping theorems. See Whitt (1974) for an interesting exposition of heavy traffic limit theorems. In the above case, the basic process turned out to be a random walk.

Instead of concentrating on a single stochastic process, Reiman (1984) considered a sequence of GI/G/1 queues indexed by  $n = 1, 2, \dots$  such that the traffic intensity  $\rho_n$  approaches to 1 in the limit. As a consequence, he obtained heavy traffic limits for various queue related processes. Using this approach we discuss in some detail about heavy traffic limit of *unfinished work process* in a GI/G/1 queue and only state the results for other queue related processes.

Consider a sequence of GI/G/1 queues on probability spaces  $\{(\Omega_n, \mathcal{E}_n, \mathcal{P}_n)\}$  with FIFO service discipline. For each  $n \geq 1$ , let  $\{u_i(n), i \geq 1\}$  and  $\{v_i(n), i \geq 1\}$  be i.i.d. sequences of positive inter-arrival times and service times respectively, with

$$\begin{aligned}\lambda_n^{-1} &= E[u_1(n)] & a_n &= \text{var}[u_1(n)] \\ \mu_n^{-1} &= E[v_1(n)] & s_n &= \text{var}[v_1(n)].\end{aligned}$$

Let

$$T_n(k) \equiv \sum_{i=1}^n u_i(n), k \geq 1, n \geq 1 \text{ and } T_n(0) = 0$$

be the arrival time of the  $k$ -th customer in the  $n$ -th system. With the inter-arrival time sequence and service time sequence, we can associate the following renewal processes respectively:

$$A_n(t) = \max\{k \geq 0 : T_n(k) \leq t\}, \quad (6)$$

$$S_n(t) = \begin{cases} 0 & \text{if } v_1(n) \geq t \\ \max\{k \geq 1 : \sum_{i=1}^k v_i(n) \leq t\} & \text{if } v_1(n) < t. \end{cases}$$

Further, let

$$L_n(t) = \sum_{i=1}^{A_n(t)} v_i(n), \quad (7)$$

$$V_n(t) = L_n(t) - t. \quad (8)$$

The unfinished work process  $U_n(t)$  is the sum of the service times of the customers in the queue and the remaining service time of the customer in service, if any. It is easy to see that

$$U_n(t) = V_n(t) - \inf_{0 \leq s \leq t} V_n(s). \quad (9)$$

Consider the following normalized processes: For  $0 \leq t \leq 1$ ,

$$\tilde{A}_n(t) = n^{-1/2}[A_n(nt) - \lambda_n nt] \quad (10)$$

$$\tilde{S}_n(t) = n^{-1/2}[S_n(nt) - \mu_n nt] \quad (11)$$

$$\tilde{V}_n(t) = n^{-1/2}[V_n(nt)] \quad (12)$$

$$\tilde{U}_n(t) = n^{-1/2}[U_n(nt)]. \quad (13)$$

In addition we need the following: Let

$$c_n = \sqrt{n}(\lambda_n - \mu_n) \quad (14)$$

$$X_n(t) = n^{-1/2} \sum_{i=1}^{[nt]} (v_i(n) - \mu_n^{-1}) \quad (15)$$

$$\alpha_n(t) = n^{-1}A_n(nt) \quad (16)$$

$$\eta_n(t) = n^{-1}S_n(nt), \text{ for } 0 \leq t \leq 1 \text{ and } n \geq 1. \quad (17)$$

Assume that

$$c_n \rightarrow c, \lambda_n \rightarrow \lambda, \mu_n \rightarrow \mu, a_n \rightarrow a, s_n \rightarrow s \text{ as } n \rightarrow \infty. \quad (18)$$

Further, assume that

$$\sup_{n \geq 1} E[(u_1(n))^{2+\epsilon}] < \infty \text{ for some } \epsilon > 0 \quad (19)$$

$$\sup_{n \geq 1} E[(v_1(n))^{2+\epsilon}] < \infty \text{ for some } \epsilon > 0 \quad (20)$$

**Theorem 2.3.1** *If (18), (19), and (20) hold, then*

$$\tilde{U}_n \Rightarrow \hat{U} \equiv RBM[c/\mu, \lambda(a+s)] \text{ in } D.$$

**Proof:** Combining the normalized processes  $\tilde{A}_n(t)$ ,  $X_n(t)$ , and  $\alpha_n(t)$ , we can write

$$\tilde{V}_n(t) = X_n \circ \alpha_n(t) + \mu^{-1}[\tilde{A}_n(t) + c_n t] \text{ for } 0 \leq t \leq 1 \text{ and } n \geq 1. \quad (21)$$

From the *functional central limit theorem for renewal processes*, it follows that  $\tilde{A}_n(t) \Rightarrow \bar{A}(t) = \text{BM}(0, \lambda^3 a)$ . Hence,  $\mu_n^{-1}[\tilde{A}_n(t) + c_n t] \Rightarrow \text{BM}[c/\mu, \lambda^3 \mu^{-2} a]$ . Now,  $\alpha_n(t) = n^{-1/2} \tilde{A}_n(t) + \lambda_n t$ . Hence,  $\alpha_n(t) \Rightarrow \lambda t$  because the first term on the RHS converges to zero functional.

Thus, from the *random time change theorem*, it follows that

$$X_n \circ \alpha_n(t) \Rightarrow \text{BM}(0, \lambda s)$$

From the asymptotic independence of the two terms on the RHS of (21), it is easy to see that

$$\tilde{V}_n(t) \Rightarrow \bar{V} = \text{BM}[c/\mu, \lambda(a+s)] \text{ in } D$$

as  $\lambda/\mu=1$  is a necessary condition for  $c$  to be finite.

If  $f : D \rightarrow D$  is defined as:

$$f(X) = X(t) - \inf_{0 \leq s \leq t} \{X(s)\}, \quad 0 \leq t \leq 1, \text{ for all } X \in D,$$

then, from the continuity of  $f$  on  $D$  in Skorohod topology, we get

$$\tilde{U}_n \Rightarrow \bar{U} = f(\bar{V}) = \text{RBM}[c/\mu, \lambda(a+s)] \text{ in } D.$$

using the *continuous mapping theorem* of §2.1.  $\square$ .

Under the assumptions (18), (19), and (20), the normalized waiting time process  $\tilde{W}_n$  and the normalized queue length process  $\tilde{Q}_n$  converge weakly to the following limits respectively:

$$\tilde{W}_n \Rightarrow \bar{U} \equiv \text{RBM}[c/\mu, \lambda(a+s)] \quad (22)$$

$$\tilde{Q}_n \Rightarrow \bar{Q} \equiv \text{RBM}[c, \lambda^3(a+s)]. \quad (23)$$

For a proof of this see Flores (1985). Reiman (1984) extended these results to queueing networks in which  $K$  GI/G/1 queues are inter-connected to form a network and the servers serve customers in FIFO order. In this case, the vector queue length process converges weakly to a  $K$ -dimensional reflected Brownian motion on the non-negative orthant  $R_+^K$ . Also results are available for sojourn time process, which is more important than queue length process in communication networks. Also results are available for the case where different types of routing and where dependencies between the arrival and service processes exist. See Reiman (1982, 1984) for details. Flores (1985) gives a survey of the results available in heavy traffic theory.

Now, suppose that we want to approximate the behaviour of the queue length process  $Q_n(\cdot)$  when the  $n$ -th system is stable but only just so. If we set  $\gamma_n = \frac{\lambda_n - \mu_n}{c}$ , then  $n^{1/2} \sim 1/\gamma_n$  for large  $n$  so that expansion of time scale by a factor  $n$  and normalization of the process  $Q_n(\cdot)$  that appeared in heavy traffic limit theorem is equivalent to expansion of time scale by a factor of  $1/\gamma_n^2$  and normalization by a factor of  $1/\gamma_n$ . Thus, we can interpret equation (23) as  $\gamma_n Q_n(\cdot/\gamma_n^2)$  converges weakly to  $\text{RBM}[c, \lambda(a+s)]$ . For each fixed  $t$ , the distribution of  $\gamma_n Q_n(t/\gamma_n^2)$  converges in distribution to  $\bar{Q}(t)$ . Hence, for large  $n$ , we might consider approximating the behaviour of  $\gamma_n Q_n(\cdot/\gamma_n^2)$ . Moreover, the limiting distribution of  $\bar{Q}(t)$  is given by,

$$\lim_{t \rightarrow \infty} \mathcal{P}\{\bar{Q}(t) \leq x\} = 1 - e^{-2|c| \cdot x/\sigma^2} \text{ for each } x \geq 0.$$

Thus, if  $Q_n(t) \Rightarrow Q_n'$  as  $t \rightarrow \infty$ , then for sufficiently large  $n$  we might approximate the distribution of  $\gamma_n Q_n'$  by the exponential distribution with parameter  $2|c|/\sigma^2$ . Diffusion approximations are based on this idea and when queueing systems are under heavy traffic such approximations yield good results. Lemoine (1978) gives a tutorial introduction to diffusion approximations.

However, when a queue is stable and if we approximate its behaviour by its limiting behaviour under heavy traffic, the effectiveness of such approximations depends on the parameters of the approximating diffusion process. Different approximating diffusions may lead to limiting diffusions with identical parameters because, in the limit, traffic intensity is equal to 1 and hence, several parameters are equal. Further, the diffusion approximations of different processes are related and thus, may lead to different approximations for the same quantity. Then these approximations need to be evaluated by their performance relative to various consistency checks. Three such approximations are given for mean in Flores (1985). These are evaluated

according to the existing upper bounds for mean delay. Whitt (1982) discusses several possible refinements to these approximations.

In the case of queueing networks, the situation is more complicated. Here also the parameters of a limiting diffusion can be written in several ways because in the heavy traffic, the arrival and service rates are equal and can be interchanged. This gives different approximations for stable systems. However, by a careful selection of the parameters of the limiting diffusion, the exact behaviour of stable queueing systems are obtained in simple cases. For example, diffusion approximation gives exact value of mean queue length for Jackson networks. See Flores (1985) for further details.

### 3. Brownian networks

As seen in the previous section, when a network of queues is under heavy traffic, i.e., when each queue is loaded to its capacity, various queueing processes converge weakly to multi-dimensional reflected Brownian motion. This is the underlying idea in the Brownian network model to be discussed in this section. The approximation involved is a system approximation; not just the approximation of one stochastic process by another. This feature gives dynamic control capability to any problem under consideration as we shall see in later sections.

In §3.1, we discuss development of Brownian network model for a multi-class open queueing network. In §3.2, a useful model in terms of workloads at stations is derived. In §3.3, we describe how the Brownian model of §3.1 can be modified to address the case of closed queueing networks. For this we need the following probabilistic setting which will be used throughout this paper.

A stochastic process will be described RCLL if its sample paths are right continuous and have left limits w.p.1. When we say  $X$  is a  $K$ -dimensional  $(\mu, \Sigma)$  Brownian motion, it is assumed that there is given a *filtered probability space*  $(\Omega, \mathcal{F}, \mathcal{F}_t, X, \mathcal{P}_x)$ , where  $(\Omega, \mathcal{F})$  is a measurable space, and  $X = X(\omega)$  is a measurable mapping of  $\Omega$  into  $C(R^K)$  which is the space of continuous functions on  $R^K$ .  $\mathcal{F}_t \equiv \sigma(X(s), s \leq t)$  is the filtration generated by  $X$  and  $\mathcal{P}_x$  is a family of probability measures on  $\Omega$  such that the process  $\{X(t), t \geq 0\}$  is a Brownian motion with drift vector  $\mu$  and covariance matrix  $\Sigma$  and initial state  $x$  under  $\mathcal{P}_x$ . Let  $E_x$  be the expectation operator associated with  $\mathcal{P}_x$ . If  $Y = \{Y(t), t \geq 0\}$  is a process that is  $\mathcal{F}_t$  measurable for all  $t \geq 0$ , then we say that the process  $Y$  is non-anticipating w.r.t  $X$  when  $Y$  is adapted to the coarsest filtration w.r.t which  $X$  is adapted. (See Harrison 1985).

Three basic notions in a *Brownian network* model are:

- *resources* indexed by  $i = 1, \dots, I$
- *activities* indexed by  $j = 1, \dots, J$
- *stocks* indexed by  $k = 1, \dots, K$

The system dynamics of a *Brownian network* can be compactly expressed by:

P.3:

$$\begin{aligned} Z(t) &= X(t) + RY(t) \in S \quad \forall t \geq 0 \\ U(t) &= AY(t) \text{ is a nondecreasing process with } U(0) = 0 \end{aligned}$$

where  $X(t)$  is a  $K$ -dimensional Brownian motion,  $R$  and  $A$  are  $K \times J$  input-output matrix and  $I \times J$  resource consumption matrix respectively. In the ensuing sections we shall see how the dynamics of a queueing network are related to that of the corresponding Brownian network.

### 3.1 Brownian approximations for scheduling multiclass open queueing networks

Consider an open queueing network with  $I$  single server stations (index  $i=1, \dots, I$ ) and with  $K$  customer classes, indexed by  $k=1, \dots, K$ . It is assumed that the class designation of a customer, summarizes all relevant and observable properties of the customer, including possibly its past processing history that may be used in dynamically scheduling the network. Customers of class  $k$  arrive according to a renewal process at an average rate of  $\lambda_k$ . It is assumed that customers of class  $k$  visit station  $s(k)$  for service and service times are i.i.d. with mean  $m_k$  and finite variance. The arrival processes and service time sequences for various classes are assumed to be mutually independent.

A customer of class  $k$  after completion of service at station  $s(k)$  will turn into a class  $j$  customer with probability  $P_{kj}$  independent of previous history. The Markovian switching matrix ( $K \times K$ ),  $P = (P_{kj})$  is assumed to be transient and hence, a customer of class  $k$  leaves the system with a positive probability  $1 - \sum_j P_{kj}$ . Let  $C(i)$  denote the constituency of server  $i$ , i.e.,

$$C(i) = \{k : s(k) = i\}, \quad i = 1 \dots I$$

From the description above, it follows that  $C(i) \cap C(j) = \phi$ ,  $i \neq j$ . As the number of classes is allowed to be arbitrary, the above routing structure is extremely general. The case where a system is populated by various customer types, each of which has an arbitrary deterministic route through the network can also be handled by assigning different class for each combination of customer type and its stage of completion. Further, Markovian switching enables to incorporate probabilistic route structure arising out of rework, spoilage, etc.

In view of the aforementioned Brownian network model (P.3), it is easy to see that queue lengths correspond to stocks, servers at  $I$  stations play the role of resources and servicing of class  $j$  customer corresponds to activity  $j$ . One unit of activity  $j$  is interpreted as one time unit allocated to class  $j$  customer by server  $s(j)$ . Activity  $j$  consumes resource  $i$  at rate,

$$A_{ij} = \begin{cases} 1 & \text{if } i = s(j) \\ 0 & \text{otherwise.} \end{cases}$$

and total amount of resource available is,  $b_i = 1$ ,  $i = 1 \dots I$ . Queue length of class  $k$  decreases by an activity  $j$  at a rate of,

$$R_{kj} = \mu_j (\delta_{jk} - P_{jk}), \text{ where } \mu_j = 1/m_j. \quad (24)$$

where  $\delta_{jk}$  is the Dirac delta function, given by

$$\delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases}$$

In the matrix form, (24) can be rewritten as:

$$R = (I - P)^T D^{-1} \quad (25)$$

where  $D$  is the diagonal matrix with elements  $m_1 \dots m_K$  and  $I$  is the  $K \times K$  identity matrix.

Since  $P$  is transient,  $R$  is non-singular with:

$$R^{-1} = D[I + P + P^2 + \dots] \quad (26)$$

Thus, there exists a unique solution  $\beta = (\beta_k)$  to:

$$\lambda - R\beta = 0 \quad (27)$$

where  $\lambda = (\lambda_k)$  is the  $K$ -vector of arrival rates. Here  $\beta_k$  can be interpreted as the average amount of time the server  $s(k)$  must assign to class  $k$  customer in order to maintain material balance over the long run.

The traffic intensity at station  $i$ ,  $\rho_i$ , is defined as

$$\rho_i = \sum_{k \in C(i)} \beta_k \quad (28)$$

Define  $\alpha = (\alpha_k)$ , the  $K$ -vector of work load proportion by:

$$\alpha_k = \frac{\beta_k}{\rho_i} \quad (29)$$

$\alpha_k$  represents the long run fraction of server's active time at station  $s(k)$  that is to be devoted to class  $k$  customer in order to maintain material balance.

Basic flow processes involved in the queueing network can be given in terms of number of customers as a  $K$ -dimensional vector process,  $F^j = \{F^j(t), t \geq 0\}$  indexed by  $j = 0, 1, \dots, K$ .  $F_k^0(t)$  is interpreted as exogenous arrival process for class  $k$  customer and  $F_k^{(j)}(t)$ ,  $j = 1, \dots, K$  is interpreted as the flow *out* of class  $k$  resulting from the  $t$  time units that server  $s(i)$  devotes to class  $j$ .

Denote by  $R^j$ , the  $j$ th column of the  $K \times K$  matrix  $R$ . Using the results of *renewal theory* (see Wolff 1989) or Karlin & Taylor (1981), it can be shown that

$$E[F^0(t)] \sim \lambda t \text{ and } E[F^j(t)] \sim R^j(t), \quad j = 1, \dots, K \quad (30)$$

Scheduling policy is expressed as a family of allocation processes,

$$T_k = \{T_k(t), t \geq 0\}, \quad k = 1, \dots, K$$

where  $T_k(t)$  gives the cumulative amount of time that server  $s(k)$  allocates to class  $k$  customers during the interval  $[0, t]$ . Then the  $K$ -dimensional queue length process  $\{Q(t), t \geq 0\}$  can be written in terms of the flow processes  $\{F^j(t), t \geq 0\}$  as follows:

$$Q(t) = F^0(t) - \sum_{j=1}^K F^j(T_j(t)). \quad (31)$$

Similarly, the  $I$ -dimensional cumulative idle time process  $\{I(t), t \geq 0\}$  can be defined by:

$$I_i(t) = t - \sum_{k \in C(i)} T_k(t), \quad i = 1, \dots, I. \quad (32)$$

The allocation process  $T = (T_k)$  reflects a scheduling policy for the queueing network and thus, we can say  $T$  is a feasible policy if it satisfies:

$$T \text{ is continuous with } T(0) = 0. \quad (33)$$

$$T \text{ is nondecreasing.} \quad (34)$$

$$T \text{ is nonanticipating with respect to } Q. \quad (35)$$

$$I \text{ is nondecreasing with } I(0) = 0. \quad (36)$$

$$Q(t) \geq 0 \quad \forall t \geq 0. \quad (37)$$

Only (35), (36), and (37) need explanation. (35) demands that scheduling policy is to be based on observable quantities. (36) expresses that a server has only  $t - s$  units of time for allocation in any interval  $[s, t]$ . (37) enforces that the server at station  $s(k)$  must stop allocating time to class  $k$  when  $Q_k(t)$  hits zero.

Thus, having expressed the basic queue processes in terms of the flow processes and the allocation process, we set out to define centred versions of these processes so as to establish connection between the system dynamics of the original queueing network and that of the approximating Brownian network, that appeared at the beginning of this section.

If we set  $T_k(t) = \alpha_k t$ , it is easy to see that such an allocation process  $T$  fully utilizes all available resources. In fact, observe that

$$A\alpha = b \quad (38)$$

Such an allocation is referred to as *nominal activity plan*. Now, for each  $k = 1, \dots, K$ , define a centred allocation process by:

$$V_k(t) \equiv \alpha_k t - T_k(t). \quad (39)$$

expressing the actual allocation to class  $k$ ,  $(T_k(t))$  as a decrement from the nominal allocation  $(\alpha_k t)$  or in vector form (39) can be written as,

$$V(t) = \alpha t - T(t). \quad (40)$$

Similarly, we can define centred flow processes as,

$$\eta^0(t) = F^0(t) - \lambda t \text{ and } \eta^j(t) = F^j(t) - R^j t, \quad j = 1, \dots, K. \quad (41)$$

Using these centred processes the queue length process can be re-expressed as

$$Q(t) = (\eta^0(t) + \lambda t) - \sum_{j=1}^K [\eta^j(T_j(t)) + R^j T(t)] \quad (42)$$

$$= (\eta^0(t) + \lambda t) - \sum_{j=1}^K [\eta^j(T_j(t)) - R T(t)]. \quad (43)$$

(43) can be compactly written as,

$$Q(t) = \zeta(t) + R V(t) \quad (44)$$



where,

$$\zeta(t) = \eta^0(t) - \sum_{j=1}^K \eta^j(T_j(t)) + (\lambda - R\alpha)t. \quad (45)$$

A similar representation for the cumulative idleness process is given as follows: observe that  $I(t) = bt - AT(t)$ . Hence it follows that

$$I(t) = AV(t). \quad (46)$$

(44) and (46) describe the system dynamics of the original queueing network and processes involved resemble the corresponding processes that appeared in the approximating Brownian network described in (P.3), with an exception that a  $K$ -dimensional Brownian motion is present instead of  $\zeta(t)$ . Thus, the essence of Brownian approximation lies in the approximation for  $\zeta$ .

Suppose that in (45), the allocation process  $T_j(t)$ , is replaced by  $\alpha_j t$  for all  $j = 1, \dots, K$ . Then, it is easy to verify that (45) reduces to,

$$\eta(t) \equiv F^0(t) - \sum_{j=1}^K F^j(\alpha_j(t)). \quad (47)$$

As the allocation process involved in (47) is the nominal one, the process  $\eta(t)$  is called the nominal queue length process.

The approximation is carried out in two steps; at the first level  $\zeta$  is approximated by  $\eta(t)$ , which in turn at the secondary level is approximated by a Brownian motion whose drift vector and covariance matrix coincide with the asymptotic drift and covariance of  $\eta(t)$ . If the original queueing network is under balanced heavy loading conditions and if the relevant processes are normalized in a manner consistent with the state of affairs, the above procedure provides good approximation.

The asymptotic drift vector  $\Upsilon$  and covariance matrix  $\Gamma$  of the process  $\eta(t)$  can be calculated using standard results of *renewal theory*. Interested reader is referred to Reiman (1984) or Harrison (1988). Thus,

$$E\{\eta(t)\} \sim \Upsilon t \text{ and } cov\{\eta(t)\} \sim \Gamma t, \text{ as } t \rightarrow \infty. \quad (48)$$

Using the *central limit theorem for random vectors* (Breiman 1968) and the *central limit theorem for renewal processes* (Wolff (1989), Karlin & Taylor (1981)), it can be shown that

$$n^{-1/2} [\eta(n) - n\Upsilon] \xrightarrow{D} N(0, \Gamma) \text{ as } n \rightarrow \infty.$$

Thus, the asymptotic distribution of  $\eta$  is the multi-variate normal distribution with mean 0 and covariance matrix  $\Gamma$ , or more generally, for each  $t > 0$  fixed,

$$\xi^*(t) \equiv n^{-1/2} [\xi nt - n\Upsilon t] \xrightarrow{D} N(0, \Gamma t) \text{ as } n \rightarrow \infty. \quad (49)$$

Thus, if  $B(t)$  is a  $K$ -dimensional Brownian motion with drift 0 and covariance matrix  $\Gamma$ , then  $\xi^*(t)$  and  $B(t)$  have approximately the same distribution for each fixed  $t$  and for large  $n$ .

Now, we will discuss the scaling operation that appeared in (49). Assume that the total work load at each station is approximately equal to its capacity in the

following sense:

there exists a large integer  $n$  such that

$$n^{1/2} |1 - \rho_i| \text{ is of moderate size for } i = 1, \dots, I. \quad (50)$$

In this case the system has balanced flow and this condition is referred to as *balanced heavy loading condition*. This  $n$  serves as an essential parameter in scaling various queueing processes. In most cases scaling expresses time as multiples of  $n$  and queue lengths as multiples of  $n^{1/2}$ . For example,  $K$ -dimensional scaled queue length process is defined by,

$$Z(t) \equiv n^{-1/2} Q(nt), \quad t \geq 0. \quad (51)$$

Similarly, the scaled versions of the processes  $\zeta$ ,  $V$ , and  $I$  are defined by,

$$X(t) \equiv n^{-1/2} \zeta(nt); \quad Y(t) \equiv n^{-1/2} V(nt) \text{ and } U(t) \equiv n^{-1/2} I(nt) \quad (52)$$

(44) and (46) can be re-expressed in terms of the above scaled processes as

$$Z(t) = X(t) + RY(t) \quad (53)$$

$$U(t) = AY(t). \quad (54)$$

and the scaled and centred allocation process  $Y(t)$  is feasible iff

$$Y \text{ is continuous with } Y(0) = 0 \quad (55)$$

$$Y(t) - Y(s) \leq n^{1/2} \alpha(t - s) \text{ if } t > s \quad (56)$$

$$Y \text{ is nonanticipating w.r.t } Z \quad (57)$$

$$U \text{ is nondecreasing with } U(0) = 0 \quad (58)$$

$$Z(t) \geq 0 \quad \forall t \geq 0 \quad (59)$$

If we define  $\theta = n^{1/2} \Upsilon$  and  $T_j^*(t) = n^{-1} (T_j(nt))$ ,  $X(t)$  can be rewritten in terms of these quantities as,

$$X(t) = n^{-1/2} [\eta^0(nt) - \sum_{j=1}^K n^{-1/2} \eta^j(n T_j^*(t)) + \theta t]. \quad (60)$$

Using the nominal activity plan for  $T_j^*(t)$ ,  $X(t)$  can be expressed in terms of centred and scaled nominal queue length process  $\xi^*(t)$  as

$$X(t) = \xi^*(t) + \theta t. \quad (61)$$

As mentioned earlier, for sufficiently large  $n$ ,  $\xi^*(t)$  can be well approximated by a  $(0, \Gamma)$  Brownian motion. Hence, from (61), it follows that  $X(t)$  can be well approximated by a  $(\theta, \Gamma)$  Brownian motion process.

Replacement of  $T_j^*(t)$  by  $\alpha_j t$  can be articulated as follows: if  $n$  is large,  $t > 0$  is moderate, from the balanced heavy loading condition (50), it follows that total server idleness at each station over the long interval  $[0, nt]$  is small compared to  $n$ , under any policy which calls for all servers to be busy whenever there is work for them to do. Hence, under such full allocation policies, the relative amounts of time that servers allocate to customers of their constituent classes must coincide with workload proportions  $(\alpha_k)$  over the long run.

Thus, in the approximating Brownian network model we can take  $X(t)$  to be the  $(\theta, \Gamma)$  Brownian motion and define  $Z$  and  $U$  in terms of  $X$  and  $Y$  as given in (53) and (54). The feasibility conditions (57)–(59) can be further simplified. With  $X$  as a Brownian motion and  $Z = X + RY$  by definition, condition (59) is equivalent to a conceptually simpler requirement that  $Y$  be non-anticipating w.r.t  $X$ . Constraints (57) and (58) are too stringent to impose. We can replace them by a weaker requirement that  $Y$  be RCLL. For a defence of this proposal, consider the constraint (58). For sufficiently large  $n$ , this constraint, which imposes a limit on the rate of increase of  $Y_k(t)$ , is loose in the sense that we can enforce rapid upward movements that closely approximate even positive jumps.

Thus, in view of the above suggested changes the approximating Brownian network takes the form:

**P.3.1.1:**

choose a  $K$ -dimensional RCLL process  $Y$  such that,

$$Z(t) = X(t) + RY(t) \quad (62)$$

$$U(t) = AY(t) \quad (63)$$

$$U \text{ is nondecreasing with } U(0) = 0 \quad (64)$$

$$Y \text{ is nonanticipating w.r.t } X \quad (65)$$

$$Y(0) = 0 \quad (66)$$

where  $X(t)$  is a  $K$ -dimensional Brownian motion.

Once the scaling parameter  $n$ , satisfying (50) is chosen, the calculation of the drift vector  $\theta$  and the covariance matrix  $\Gamma$  for  $X(t)$  entails knowledge of only the first and second moments of the arrival and service patterns. Hence, the approximating Brownian network is insensitive to the specific form of the arrival and service distributions.

The decision problem in the Brownian network can be transformed into a more intuitively appealing workload problem which, besides being amenable to analytical tractability, has the advantage that given a performance measure the optimal solution is easier to interpret than that obtained by solving the original problem.

### 3.2 Workload formulation for a Brownian network problem

Define an  $I \times K$  matrix  $M = (M_{ik})$  by

$$M \equiv AR^{-1} = AD[I + P + P^2 + \dots] \quad (67)$$

$M_{ik}$  represents the expected total time that server  $i$  must allocate to a class  $k$  customer before it eventually leaves the system. Define an  $I$ -dimensional *workload* process  $W = (W_i)$  as

$$W \equiv MZ(t). \quad (68)$$

$W_i(t)$  gives the expected total amount of work embodied in those customers present anywhere in the network at time  $t$ . (Recall that all the processes in (68) are expressed in scaled units). The state space  $S$  of  $W$  is,

$$S = \{w \in R^I : w = MZ, Z \in R_+^K\}$$

Define an  $I$ -dimensional Brownian motion  $B$  as,

$$B(t) = M X(t). \quad (69)$$

The drift vector and covariance matrix of  $B(t)$  are  $M\theta$  and  $M\Gamma M^T$  respectively. With the above modifications, the decision maker's problem can be redefined as:

**P.3.2.1:**

Choose a pair of RCLL processes  $(Z, U)$  such that

$$U \text{ is nonanticipating w.r.t } B \quad (70)$$

$$U \text{ is nondecreasing with } U(0) = 0 \quad (71)$$

$$Z(t) \geq 0 \quad \forall t \geq 0 \quad (72)$$

$$M Z(t) = B(t) + U(t) \quad \forall t \geq 0. \quad (73)$$

The allocation process  $Y(t)$  can be expressed in terms of  $(Z, U)$  as,

$$Y(t) = R^{-1} [Z(t) - X(t)]$$

Equivalence of the two formulations (P.3.1.1) and (P.3.2.1) follows from the fact that  $Y(t)$  given above satisfies all the conditions given in (P.3.1.1).

### 3.3 The case of closed queueing networks

In this subsection we discuss how the approximating Brownian network described in §3.1 can be modified to address the case of closed queueing networks. In a closed queueing network, a constant population of customers circulates indefinitely through the network, with no exogenous arrivals and departures. An initial queue length vector  $Q(0)$  is specified a priori. Further, the switching matrix  $P$  in this case is irreducible and hence, is of rank  $K-1$ , which further implies that the input-output matrix is of rank  $K-1$ .

Analogous to the case discussed in §3.1, we seek a  $K$ -vector  $(\beta)$  of average activity rates satisfying,

$$R\beta = 0 \quad (74)$$

Equation (74) has strictly positive solution unique only up to a scale constant. Thus the traffic intensities,

$$\rho_i = \sum_{k \in C(i)} \beta_k, \quad i = 1, \dots, I. \quad (75)$$

are determined up to a scale constant. To resolve this ambiguity, the average activity rates  $(\beta_1, \dots, \beta_K)$  are scaled so that  $\max_i \rho_i = 1$ . In this case the traffic intensities  $\rho_i$  express the relative amounts of work that servers at the various stations must do to maintain material balance. If we call the station  $k$  with  $\rho_k = 1$  as bottleneck station, then  $\rho_i$  represents the fraction of time that server  $i$  would be kept busy if the bottleneck station is never idle.

The analog of heavy traffic condition (50) in this case is that:

There exists a large integer  $n$  such that

$$n^{1/2} |1 - \rho_i| \text{ is of moderate size} \quad (76)$$

$$n^{-1/2} [Q_1(0) + \dots + Q_K(0)] = 1 \quad (77)$$

In other words, the total population size  $N$  in the network should be such that  $|\rho_i - \rho_j|$  for  $i \neq j$ , is of order  $N^{-1}$  or smaller for each pair of  $i$  and  $j$  and that we choose  $n = N^2$  as scaling parameter for the approximating Brownian network.

Following the notation of the §3.1, the vector queue length process is given by,

$$Q(t) = Q(0) - \sum_{j=1}^K F^j(T_j(t)). \quad (78)$$

The components of  $F^j(t)$  sum to zero so that the queue length remains constant over time.

Nominal allocation for class  $k$  over  $[0, t]$  can be taken to be  $\alpha_k t$ , as in the earlier case, where,

$$\alpha_k = \frac{\beta_k}{\rho_i} \quad \forall k \in C(i) \quad (79)$$

Embedding the initial queue length vector  $Q(0)$  in the definition of  $\zeta(t)$ , we get

$$\zeta(t) = Q(0) - \sum_{j=1}^K \eta^j(T_j(t)) - R\alpha t. \quad (80)$$

Then the identity (44) remains valid in the closed network case. Similarly, incorporation of  $Q(0)$  in the definition of  $\xi(t)$  gives the nominal queue length process as

$$\xi(t) = Q(0) - \sum_{j=1}^K F^j(\alpha_j t). \quad (81)$$

The asymptotic drift vector and covariance matrix of  $\xi(t)$  satisfy,

$$e^T \Upsilon = 0 \quad \text{and} \quad e^T \Gamma e = 0, \text{ where } e \text{ is a } K - \text{dimensional sum vector.}$$

Justification for using nominal allocation in the approximating Brownian network, in this case, can be given as follows: in closed queueing networks, the decision maker's problem is to maximize rate of circulation, which boils down to maximizing the fraction of the time that any server is kept busy. Hence, full allocation policy is justified and the approximation is valid under any such policy.

A few more changes need be taken into account in the case of closed networks. The underlying Brownian motion  $X(t)$  has now the initial state :

$$X(0) = Z(0) = n^{-1/2} Q(0). \quad (82)$$

It is easy to see that  $e^T X(0) = e^T X(t) = e^T Z(t) = 1$ , consistent with constant population size.  $Z_k(t)$ , hence, can be interpreted as *the fraction of the total population* that belongs to class  $k$  at time  $t$ .

Workload formulation given in §3.2 cannot be extended to the closed network case because here  $R$  is singular. But, using a modeling artifice, a similar transformation can be achieved as we shall see in §4.2.

#### 4. Methodology and numerical results for three different queueing systems

##### 4.1 Scheduling a multiclass make-to-stock queue

In a make to stock production system, products are made according to a forecast of demand and completed jobs enter a finished good inventory which services actual customer demand. Here, we consider a simple case of *make-to-stock* system with a single machine centre.  $K$  classes of products are made and service times of products of class  $k$  have a general distribution with mean  $m_k$  and finite squared co-efficient of variation,  $v_{s_k}^2$ . Demand for products of class  $k$  is a renewal process with rate  $\lambda_k$  and squared coefficient of variation,  $v_{d_k}^2$ . Holding cost of  $h_k$  units per unit time is incurred for maintaining inventory of class  $k$  products and a back order cost of  $b_k$  units per unit time is incurred if inventory of class  $k$  is not available.

It is assumed that ample amount of raw material is available for all types of products and also that no set up time/cost is incurred when the machine switches over from one class to another. The scheduling problem is to choose among  $K+1$  options, i.e.,

- either work on a class  $k$  job,  $k = 1, \dots, K$
- or allow the machine to idle.

with a view to minimize the long run expected cost incurred.

Let  $\{S_k(t), t \geq 0\}$  be the renewal process associated with the service times of class  $k$ , giving at any point of time  $t$ , number of service completions in the interval  $[0, t]$ . Let  $\{D_k(t), t \geq 0\}$  be the point process for demands which gives number of class  $k$  demands up to time  $t$ . The inventory level process  $Z_k(t)$  is given by,

$$Z_k(t) \equiv S_k(T_k(t)) - D_k(t), \quad (83)$$

where  $T_k(t)$  is the allocation process which at time  $t$  gives the cumulative amount of time allotted to class  $k$  in the interval  $[0, t]$ . Thus, the decision maker's problem is to

(P.4.1.1)

choose a  $K$ -dimensional allocation policy  $T = (T_k)$  to

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E \left[ \int_0^T \sum_{k=1}^K c_k(Z_k(t)) dt \right]$$

where,

$$c_k(x) = \begin{cases} h_k x & \text{if } x \geq 0 \\ -b_k x & \text{if } x < 0. \end{cases}$$

subject to

$$T \text{ is nondecreasing and continuous with } T(0) = 0 \quad (84)$$

$$T \text{ is nonanticipating w.r.t } Z \quad (85)$$

$$I \text{ is nondecreasing with } I(0) = 0 \quad (86)$$

To develop the Brownian approximation for the problem (P.4.1.1), we consider centred and scaled versions of all the related processes.

Define the traffic intensity of the system by,

$$\rho = \sum_{k=1}^K \rho_k \quad \text{where } \rho_k = \frac{\lambda_k}{\mu_k}. \quad (87)$$

$\rho$  gives average server utilization required to satisfy the average demand.

Define  $\alpha_k = \rho_k/\rho$  to be the proportion of the server's busy time that should be devoted to class  $k$  to meet average demand. The centred allocation process and the centred renewal process generated by service completions are given respectively as:

$$Y_k(t) = \alpha t - T_k(t) \quad (88)$$

$$\eta_k(t) = S_k(t) - \mu_k t \quad (89)$$

Furthermore define,

$$\Psi_k(t) \equiv (\mu_k \alpha_k - \lambda_k)t + \eta_k(T_k(t) - D_k(t) + \lambda_k t) \quad \text{for } k = 1, \dots, K \text{ and } t \geq 0 \quad (90)$$

(Note that  $\Psi(t)$  corresponds to the process  $\zeta(t)$  of (45)). Then the queue length process and idle time process can be reexpressed in terms of (88) and (89) as follows:

$$Z_k(t) = \Psi_k(t) - \mu_k Y_k(t) \quad \forall k = 1, \dots, K \text{ and } t \geq 0. \quad (91)$$

$$I_k(t) = \sum_{k=1}^K Y_k(t) \quad \forall t \geq 0 \quad (92)$$

Now, choosing the scaling parameter as  $(1-\rho)^2$ , the above basic processes will be normalized as given below. (For notational convenience, the same symbols are used for scaled processes).

$$Z_k(t) = \frac{Z_k(nt)}{\sqrt{n}}, \quad k = 1, \dots, K \quad \forall t \geq 0 \quad (93)$$

$$Y_k(t) = \frac{Y_k(nt)}{\sqrt{n}}, \quad k = 1, \dots, K \quad \forall t \geq 0 \quad (94)$$

$$I(t) = \frac{I(nt)}{\sqrt{n}}, \quad \forall t \geq 0 \quad (95)$$

We get the nominal inventory level process by replacing  $T_k(t)$  in (90) by  $\alpha_k t$ . Then, using the central limit theorem for renewal processes (see Wolff 1989), the random time change theorem and the continuous mapping theorem discussed in §(2.1), we can show that the nominal inventory level process  $\Psi_k(t)$  converges weakly to a Brownian motion  $X_k(t)$  with drift  $\sqrt{n}(\lambda_k - \mu_k \alpha_k)$  and variance  $\lambda_k (v_{s_k}^2 + v_{d_k}^2)$ .

Thus, the approximating Brownian control problem for (P.4.1.1) appears as follows:

Choose a policy ( $Y$ ) to

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T \sum_{k=1}^K c_k(Z_k(t)) dt \right]$$

subject to

$$Z_k(t) = X_k(t) - Y_k(t) \text{ for } k = 1, \dots, K, \forall t \geq 0 \quad (96)$$

$$I(t) = \sum_{k=1}^K Y_k(t) \quad (97)$$

$$I \text{ is nondecreasing with } I(0) = 0 \quad (98)$$

$$Y \text{ is nonanticipating w.r.t } X \text{ and } Y(0) = 0 \quad (99)$$

#### Workload formulation:

The workload process  $W(t)$ , which gives at any time  $t$  the expected amount of total work embodied in the system, is given by,

$$W(t) = \sum_{k=1}^K m_k Z_k(t). \quad (100)$$

Define the one dimensional Brownian motion  $B$  by

$$B(t) = \sum_{k=1}^K m_k X_k(t), \forall t \geq 0,$$

so that  $B$  has drift  $\delta = \sqrt{n}(1 - \rho) > 0$  and variance  $\sum_{k=1}^K \lambda_k m_k^2 (v_{s_k}^2 + v_{d_k}^2)$ . Then, the workload formulation for the problem (P.4.1.2) is:

#### P.4.1.3:

choose the pair  $(Z, I)$  so as to

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T \sum_{k=1}^K c_k(Z_k(t)) dt \right]$$

subject to

$$W(t) = B(t) - I(t), \forall t \geq 0 \quad (101)$$

$$I \text{ is nondecreasing with } I(0) = 0 \quad (102)$$

$$Z \text{ and } I \text{ are nonanticipating w.r.t } B \quad (103)$$

The equivalence of (P.4.1.2) and (P.4.1.3) can be established easily (see Wein 1992b).

The problem (P.4.1.3) is easier to solve and its solution easier to interpret in terms of the original queueing system, compared to problem (P.4.1.2). We briefly sketch the solution procedure for (P.4.1.3) and urge the reader to see Wein (1992) for further details.

Observe that given  $I(t)$  at each point of time  $t$ , which satisfies the constraints (102) and (103), embedded in the problem (P.4.1.3) is a *linear programming problem*. A simple closed form solution can be obtained in terms of  $W(t)$  by reformulating the



problem as a linear programming problem with separate variables for the positive and negative parts of  $Z_k(t)$ .

Define the indices  $j$  and  $l$  by

$$\min_{1 \leq k \leq K} \frac{h_k}{m_k} = \frac{h_j}{m_j} \quad (104)$$

$$\min_{1 \leq k \leq K} \frac{b_k}{m_k} = \frac{b_l}{m_l}. \quad (105)$$

Because the problem for a given  $I(t)$  has only one constraint, it is easy to see that the optimal solution for the linear programming problem is:

$$Z_k^*(t) = \begin{cases} \frac{W(t)}{m_k} & \text{if } k = j \text{ and } W(t) \geq 0, \\ 0 & \text{if } k \neq j \text{ and } W(t) \geq 0. \end{cases}$$

$$Z_k^*(t) = \begin{cases} \frac{W(t)}{m_k} & \text{if } k = l \text{ and } W(t) < 0 \\ 0 & \text{if } k \neq l \text{ and } W(t) < 0. \end{cases}$$

Hence, the optimal solution to (P.4.1.3) is dependent on  $I(t)$  through  $W(t)$ . Thus, the work load formulation reduces to choosing an optimal policy  $I(t)$ , which should be an RCLL process and non-anticipating w.r.t  $B$ . So the resulting Brownian control problem is to find such an  $I(t)$  to

(P.4.1.4):

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E \left[ \int_0^T f(W(t)) dt \right]$$

subject to

$$W(t) = B(t) - I(t) \quad \forall t \geq 0 \quad (106)$$

where,

$$f(x) = \begin{cases} \frac{h_j x}{m_j} & , \text{if } x \geq 0 \\ \frac{b_l x}{m_l} & , \text{if } x < 0. \end{cases}$$

Observe that from the positive drift of  $B(t)$  and the nature of  $f(t)$ , it is natural to consider a policy  $I(t)$  which keeps  $W(t)$  in an interval of the form  $[-\infty, c]$  while exerting minimum amount of control  $I(t)$ . The process  $W(t)$  under such a policy is called *regulated Brownian motion* on  $[-\infty, c]$ . A candidate policy  $I(t)$ , given by

$$I(t) = \sup_{0 \leq s \leq t} [B(s) - c]^+, \quad \forall t \geq 0. \quad (107)$$

satisfies all the requirements specified above [see Chapter 1 of Harrison (1985)]. Thus, if we confine ourselves to the policies of the type (107), the cost function appearing in (P.4.1.4) can be expressed as a function of  $c$ . For this, we need the following proposition from Harrison (1985).

**Theorem 4.1.1** Suppose that  $B$  is a  $(\mu, \sigma^2)$  Brownian motion,  $I$  is as defined in (107) and thus,  $W = B - I$  is an RBM on  $[-\infty, c]$ . Then,  $W$  has exponential steady state distribution with density,

$$p(x) = \begin{cases} \nu e^{\nu(x-c)} & \text{if } x \leq c \\ 0 & \text{if } x > c \end{cases}$$

Furthermore, for each starting state  $x < c$ , there exists a constant  $C$  such that

$$E_x[W^2(t)] < C, \quad \forall t \geq 0$$

Using the above proposition, the cost function in (P.4.1.4) can be written as

$$F(c) = - \int_{-\infty}^0 b x \nu e^{\nu(x-c)} dx + \int_0^c h x \nu e^{\nu(x-c)} dx. \quad (108)$$

and the value of  $c$  that minimizes  $F(c)$  is

$$c^* = \frac{\sigma^2}{2\mu} \ln\left(1 + \frac{b}{h}\right)$$

with  $F(c^*) = \frac{h\sigma^2}{2\mu} \ln\left(1 + \frac{b}{h}\right)$ . For the proof of optimality of the policy  $I^* = \sup_{0 \leq s \leq t} [B(s) - c^*]$ , see Wein (1992a).

Eventhough the underlying processes in deriving the optimal policy  $I^*$  are scaled versions of the corresponding processes of the original problem, still the solution can provide insights to develop an effective scheduling policy.

Recall that  $I(t)$  represents the scaled cumulative idleness process and under the optimal policy  $I^*(t)$ , the scaled and weighted inventory process  $W(t)$  is an RBM. The process  $I^*(t)$  increases only when  $W(t)$  is equal to  $c^*$  or in otherwords the server is idle only at times  $t$  when  $W(t)$  is  $c^*$  and otherwise is busy. Let  $w(t)$  be the actual unscaled weighted inventory process. Then  $W(t)$  and  $w(t)$  are related by,

$$W(t) = \frac{w(nt)}{\sqrt{n}}, \quad t \geq 0.$$

Thus, the machine should be kept busy whenever  $w(t) < \sqrt{n} c$  or when

$$w(t) < \frac{\sum_{k=1}^K \lambda_k m_k^2 (v_{s_k}^2 + v_{d_k}^2)}{2(1-\rho)} \ln\left(1 + \frac{b}{h}\right).$$

In a similar fashion, the priority scheduling decision can be in terms of the optimal inventory level process  $Z^*$ . Whenever  $w(t) < 0$ , only one component of the inventory is seen to be at the positive level. In particular, no inventory is held and back orders are all of the class with the minimum value of the index  $b_k \mu_k$  and hence the backordered demands of this class should only be satisfied when this class is the only one that is backordered at time  $t$ . In heavy traffic, the scaled number of backorders of other classes will be negligible and it does not matter in which order these backorders are satisfied. To resolve this ambiguity in priority assignment for these classes, an intuitively appealing decision would be to give priority to the class with the largest value of the index  $b_k \mu_k$  among all the classes that are back ordered at time  $t$ .

Table 1. Data for the model in figure 1

Class	Int. arr. time distribution (mean, [std.dev])	Service distribution (mean, [std.dev])	Backorder cost	Holding cost
1	Uniform(24.0, 11.54)	Uniform(2.0, 1.0)	2.0	2.0
2	Exp(150)	Exp(150)	10.0	10.0
3	Exp(60.0)	Uniform(10.0, 2.0)	100.0	5.0
4	Uniform(20.8, 8.66)	Exp(5.0)	5.0	10.0
5	Exp(60.0)	Normal(15.0, 4.0)	5.0	5.0

Extending the same arguments to the case when  $w(t) > 0$ , an effective scheduling policy is to process the class with the minimum value of  $h_k \mu_k$  whenever no jobs are backordered.

However, the foregiven scheduling policy has a shortcoming that it does not anticipate backorder job classes and does not respond to the class until its inventory level is negative. To compensate for this, a parametric policy is suggested in terms of parameters,  $\epsilon_k$ ,  $k = 1, \dots, K$  which at any time  $t$  indicate which of the classes are in *danger of being backordered*. A class  $k$  is in danger of being backordered if,  $Z_k(t) < \epsilon_k$  at any time  $t$ . In terms of the parameters  $\epsilon_k$ , the above scheduling policy can be modified as follows: the machine is idle whenever the weighted inventory process  $w(t) \geq \sqrt{n}c$  and no classes are in danger of being backordered; otherwise, is busy. Among the subset of the classes that are in danger of being backordered, priority is based on the value of the index  $b_k \mu_k$  and the class with minimum value of this index is served first. When no class is in danger of being backordered, the machine processes the classes based on the index  $h_k \mu_k$  serving the class with the minimum value of  $h_k \mu_k$  first. For a detailed description of these policies see Wein (1992b) and Veatch & Wein (1992).

#### 4.1.1 An example: A five class make-to-stock system

A simulation study is performed on the example described in §1.4. The processing time distributions, the customer inter-arrival distributions and backordering and holding costs for all the five classes are shown in table 1. High priority for class 3 is taken into account by assigning high backordering cost and low holding cost. The BROWNIAN policy described above is compared against other scheduling policies such as FCFS (First Come First Served), MIN (MINimum inventory level), SEPT (Shortest Expected Processing Time) policies. In all these policies the Busy/Idle decision is according to an dependent (S-1,S) policy for each class. Under this policy, an arriving customer simultaneously takes a class  $k$  product from the inventory (if not available, backorders for one) and initiates request for a class  $k$  product. The machine centre is busy only when requests are queued. Safety stock levels for FCFS and MIN policies correspond to the optimal stock levels obtained by performing a Brownian analysis similar to that given above. In the case of SEPT policy, the safety stock levels are arbitrarily selected using some intuitive arguments.

For all the above policies, safety stock levels and total cost achieved for that safety stock are given in table 2. It can be easily observed that BROWNIAN policy outperforms all the other policies.

Another simulation experiment is conducted for different utilization levels of the

Table 2. Costs for various policies.

Policy	Safety Stock	Avg. Cost
FCFS	(6, 1, 11, 4, 2)	772.2
MIN	(6, 1, 11, 4, 2)	676.8
SEPT	(0, 0, 3, 0, 0)	527.4
BROWNIAN	(0, 0, 2, 0, 0)	432.9

machine center and the costs incurred for all the above policies are presented in table 3. Referring to the results, it is seen that under light load conditions the Brownian policy is not as effective as the other policies. This behaviour is due to the fact that under low utilizations, the BROWNIAN policy tries to keep the machine center busy even though the inventory levels exceed the safety stock levels and the arrival rate of demands is very low. But at higher utilizations, the BROWNIAN policy dominates the other policies.

#### 4.2 Scheduling a two-station closed queueing network

Here we consider the problem of optimally scheduling a two station closed queueing network with  $K$  customer classes to maximize the long run expected average throughput of the network. We describe a Brownian model for the problem under the setting given in §§3.1 and 3.4 and follow the same notation given there.

An approximating Brownian network is developed along the same lines as described earlier except for a change in the scaling phenomenon. Here scaling re-expresses time as multiples of  $N^2$  and queue length as multiples of  $N$ , where  $N$  is the total population size, *i.e.*,

$$Z_k(t) = \frac{Q_k(N^2 t)}{N} \quad \forall k = 1, \dots, K \quad (109)$$

$$U_i(t) = I_i(t) = \frac{I_i(N^2 t)}{N} \quad \text{for } i = 1, 2, \forall t \geq 0 \quad (110)$$

The allocation process  $T(t)$  is centred by the vector  $\alpha = (\alpha_k)$  of workload proportions and then scaled to give,

$$Y_k(t) = \frac{\alpha_k N^2 t - T_k(N^2 t)}{N}, \quad \forall k = 1, \dots, K \quad (111)$$

$\alpha_k$  is as given in (79). Recall that in closed queueing networks  $Z_k(t)$  gives the fraction of the total population that belongs to class  $k$  at any time  $t$ .

In a closed queueing network, maximizing the long run average throughput rate is equivalent to minimizing the long run average amount of idleness at either station. Without loss of generality, here we seek to minimize  $U_1$ .

Thus, the Brownian control problem is to,

(P.4.2.1):

Choose a policy  $(Y)$  to

$$\text{minimize} \quad \limsup_{T \rightarrow \infty} \frac{1}{T} E[U_1(T)]$$

Table 3. Costs at different utilizations.

Utilization	FCFS	MIN	SEPT	BROWNIAN
0.03	282.2	282.1	282.3	1505.4
0.20	289.6	288.4	288.8	662.4
0.90	499.4	437.2	348.5	303.9
0.99	5365.5	3162.5	1144.7	921.1

subject to

$$Y \text{ is nonanticipating w.r.t } X \quad (112)$$

$$Z(t) = X(t) + RY(t), \forall t \geq 0 \quad (113)$$

$$U(t) = AY(t), \forall t \geq 0 \quad (114)$$

$$U \text{ is nondecreasing with } U(0) = 0 \quad (115)$$

$$e^T Z(t) = 1, \forall t \geq 0 \quad (116)$$

$$Z(t) \geq 0. \quad (117)$$

The drift  $\delta$  and covariance matrix  $\Sigma$  of the Brownian motion  $X$  are

$$\delta = -N R \alpha \quad (118)$$

$$\Sigma = \sum_{k=1}^K [\alpha_k m_k^{-1} P_{kj} (\delta_{jl} - P_{kl}) + \alpha_k m_k^{-1} s_k^2 R_{jk} R_{lk}] \quad (119)$$

where  $R$  is the input-output matrix and  $P$  is the switching matrix given in §3.1. Assume that  $P$  is irreducible. As mentioned in §3.4, the above problem defies reformulation in terms of workloads for in this case, the matrix  $R$  is singular. However, the following modeling artifice obviates this difficulty.

As it is assumed that each type has its own deterministic route through the network, each class,  $k = 1, \dots, K$ , corresponds to a particular stage along a type's route. Denote all the classes that correspond to the last stage along the route of some customer type as potential exit classes. Arbitrarily select a potential exit class, say  $K$ . Let  $q = (q_k)$  be the  $K$ th column of  $P^T$ . Thus, the elements of  $q$  give the probability of transitions from the potential exit class  $K$  and  $q_k$  is positive only for classes that correspond to the first stage along some customer type's route.

Define  $K \times K$  matrix  $\Delta$  as:

$$\Delta^j = (P^T)^j \text{ for } j = 1, \dots, K-1 \quad (120)$$

$$\Delta^K = (0) \quad (121)$$

where  $B^j$  denotes the  $j$ th column of matrix  $B$  and  $(0)$  is the  $K$ -dimensional vector of zeroes.

Since  $P$  is irreducible,  $(I - \Delta)^{-1}$  exists. Let  $D$  be the diagonal matrix with diagonal elements  $m_1, \dots, m_K$ . Define the matrix  $H$  by,

$$H = D(I - \Delta)^{-1}.$$

Now, we can define workload profile matrix,  $\overline{M}$  for closed networks as,

$$\overline{M} = A H. \quad (122)$$

$\overline{M}_{ik}$  represents the expected total time the server  $i$  devotes to a class  $k$  customer until that customer next completes service as a class  $K$  customer, i.e., until that customer *exits*.

Now, define the two dimensional scaled workload process  $W(t)$  by

$$W(t) \equiv \overline{M} Z(t) \quad (123)$$

$W_i(t)$  at any time  $t$  gives the expected total amount of work for server  $i$  embodied in all customers in the network at time  $t$  until they next complete service as a class  $K$  customer. Define the two dimensional Brownian motion ( $B$ ) by,

$$B(t) = \overline{M} X(t), \quad \forall t \geq 0 \quad (124)$$

$B$  has drift  $\overline{M} \delta$  and covariance matrix  $\overline{M} \Sigma \overline{M}^T$ .

However, in order to calculate the actual workload at any time  $t$ , we have to take into account the expected total time,  $v_i$ , that server  $i$  must devote to a class  $k$  customer until he next exits. It is easy to see that

$$v = \overline{M} q. \quad (125)$$

Average number of such *newly exiting* customers,  $\theta(t)$ , is given by,

$$\theta(t) = m_k^{-1} Y_k(t) \quad (126)$$

Then, the workload formulation for the problem (P.4.2.1) is given as follows: (P.4.2.2):

choose RCLL processes ( $Z, U, \theta$ ) to

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E [U_i(T)]$$

subject to

$$Z, U, \theta \text{ are nonanticipating w.r.t } X \quad (127)$$

$$M Z(t) = B(t) + U(t) - v \theta(t), \quad \forall t \geq 0 \quad (128)$$

$$U \text{ is nondecreasing with } U(0) = 0 \quad (129)$$

$$e^T Z(t) = 1 \quad \forall t \geq 0 \quad (130)$$

$$Z(t) \geq 0 \quad \forall t \geq 0 \quad (131)$$

For a proof of equivalence of the formulations (P.4.2.1) and (P.4.2.2) see Harrison & Wein (1990).

Interestingly, it turns out that the vector of traffic intensities  $\rho$  is proportional to the vector  $v$ , i.e.,  $\rho_i = c v_i$ . This observation facilitates further reduction in dimensionality of the problem (P.4.2.2). To see this, define the one-dimensional workload imbalance process  $\widehat{W}$  by

$$\widehat{W} = \rho_2 W_1(t) - \rho_1 W_2(t), \quad \forall t \geq 0 \quad (132)$$

If  $\widehat{W} > 0$ , then the workload in the system is imbalanced towards station 1. Define the one-dimensional Brownian motion  $\widehat{B}$  by,

$$\widehat{W} = \rho_2 B_1(t) - \rho_1 B_2(t), \quad \forall t \geq 0 \quad (133)$$

1,  $\hat{B}$  has drift  $\mu = \rho^T \bar{M} \rho$  and variance  $\sigma^2 = \rho^T \bar{M} \Sigma \bar{M}^T \rho$ , where  $\rho = \begin{pmatrix} \rho_2 \\ -\rho_1 \end{pmatrix}$ .  
easy to prove that  $\mu = N(\rho_2 - \rho_1)$ . Define one dimensional processes  $R$  and  $L$ ,

$$R = \rho_2 U_1(t), \quad \forall t \geq 0 \quad (134)$$

$$L = \rho_1 U_1(t), \quad \forall t \geq 0. \quad (135)$$

and  $L$  can be interpreted as right and left movements exerted by the controller.

Using the fact that  $\rho = cv$ , the workload problem ( P.4.2.2 ) can be reformulated as a single dimensional problem as follows:

**1.2.3):**

Choose a pair  $(R, L)$  to

$$\text{minimize} \quad \limsup_{T \rightarrow \infty} \frac{1}{T} E \left[ \frac{R(t)}{\rho_2} \right]$$

subject to

$$R \text{ and } L \text{ are nonanticipating w.r.t } \hat{B} \quad (136)$$

$$\hat{W}(t) = \hat{B}(t) + R(t) - L(t), \quad \forall t \geq 0 \quad (137)$$

$$R \text{ and } L \text{ are nondecreasing with } R(0) = L(0) = 0 \quad (138)$$

$$\hat{W}(t) = \sum_{k=1}^K (\rho_2 \bar{M}_{1k} - \rho_1 \bar{M}_{2k}) Z_k(t) \quad (139)$$

$$e^T Z(t) = 1 \quad \forall t \geq 0 \quad (140)$$

$$Z(t) \geq 0 \quad \forall t \geq 0. \quad (141)$$

A pathwise solution which minimizes  $R(t)$  and  $L(t)$  for all times  $t$  simultaneously can be found by initially ignoring the process  $Z(t)$  and replacing the constraints (136)–(141) of the problem ( P.4.2.3 ) by a surrogate condition that the process  $\hat{W}(t)$  be confined to an interval  $[a, b]$ . In view of the constraints (139), and (140), it is easy to see that the interval end points  $a$  and  $b$  are given respectively by,

$$a = \rho_2 \bar{M}_{12} - \rho_1 \bar{M}_{22} \equiv \min_k (\rho_2 \bar{M}_{1k} - \rho_1 \bar{M}_{2k}) \quad (142)$$

$$b = \rho_2 \bar{M}_{11} - \rho_1 \bar{M}_{12} \equiv \max_k (\rho_2 \bar{M}_{1k} - \rho_1 \bar{M}_{2k}) \quad (143)$$

It follows that  $a \leq 0 \leq b$ , and class 1 customers are served at station 1 and class 2 at station 2.

The pair of RCCL processes  $(R, L)$  are feasible policies only if the associated process  $\hat{W}(t)$  is kept within  $[a, b]$ . Among all the feasible policies  $(R, L)$ , the policies given by

$$R(t) = \sup_{0 \leq s \leq t} [a - \hat{B}(s) + L(s)]^+ \quad (144)$$

$$L(t) = \sup_{0 \leq s \leq t} [\hat{B}(s) + R(s) - b]^+ \quad (145)$$

minimize the values of  $R(t)$  and  $L(t)$  for all  $t$  simultaneously w.p.1. (for a proof, see Chap. 2 of Harrison (1985))

For the policies defined by (144) and (145),  $R$  and  $L$  increase only when  $\widehat{W}(t) = a$  and  $\widehat{W}(t) = b$  respectively. To find out a control process  $Z(t)$ , that completes the pathwise solution, define,

$$\gamma(t) = \frac{\widehat{W}(t) - a}{b - a} \quad \forall t \geq 0. \quad (146)$$

Let  $Z(t)$  be defined by,

$$Z(t) = \begin{cases} \gamma(t) & , \text{ if } k = 1 \\ 1 - \gamma(t) & , \text{ if } k = 2 \\ 0 & , \text{ otherwise.} \end{cases}$$

Thus,  $(Z, R, L)$  defined by (144) and (145) respectively give a pathwise solution to the problem (P.4.2.3). Proof of this can be found in Harrison & Wein (1990). The solution  $(Z, U, \theta)$  for (P.4.2.2) can be found from the existing relations.

From the solution for  $Z(t)$ , it follows that in heavy traffic limit only two components indexed by 1 and 2 are positive. Class 1 is served at station 1 and class 2 at station 2. This solution can be interpreted to mean that the classes 1 and 2 are to be given lowest priority at the respective stations. Under the heavy traffic condition, it does not matter in whatever order the other  $K-2$  classes are served. However, to be specific, a natural ordering based on workload imbalance indices minimizes the idle time of any server. To see this, observe that idleness is incurred only when  $\widehat{W}(t) = a$  or  $b$ . Order the classes now according to the values

$$\rho_2 \overline{M}_{1k} - \rho_1 \overline{M}_{2k}. \quad (147)$$

Suppose that priority rule assigns highest priority at station 1 (respectively, at station 2) to the classes with smaller (respectively, larger) values of the index. Then, the workload imbalance process  $\widehat{W}(t)$  is kept within the interval  $[a, b]$ . As a result, idleness will be incurred less often than any other sequencing policy, such as SPT, SRPT etc. For a formal justification of this fact, see Harrison & Wein (1990). The foregoing scheduling problem in multi-class case was discussed by Chevalier & Wein (1993).

#### 4.2.1 An example: A closed re-entrant line

We now present the report of simulation studies performed on the re-entrant line example of §1.4. The service time distributions for all the classes are given in table 4. In this example, machine centers 1 and 2 act as bottlenecks. The simulation experiment is conducted for different population sizes where WBAL (Workload BALancing) scheduling policy is followed at stations 1 and 2 and FCFS is followed at the machine center 3. Also the experiment is performed with other conventional scheduling policies which include FCFS, SEPT, LBFS (Last Buffer First Served) and FBFS (First Buffer First Served) policies. WBAL policy awards priority, from high to low, to classes (1,9,8) at station 1 and (7,4,10,3,2) at station 2 whereas the priority order for SEPT policy is easily seen to be (9,8,1), (2,7,10,4,3) and (5,11,6)



**Table 4.** Data for the model in figure 2.

Class	Service Distribution (mean, [std.dev])
1	Exp(9.0)
2	Uniform(1.0, 0.25)
3	Exp(8.0)
4	Exp(6)
5	Uniform(2.0, 0.7)
6	Normal(6.0, 1.0)
7	Uniform(3.0, 0.7)
8	Exp(8.0)
9	Normal(6.0, 1.0)
10	Exp(5.0)
11	Exp(5.0)

ations one, two, three respectively. Mean cycle times and variances of cycle s for a given throughput rate are then compared in table 5.

ie reason for comparing under constant throughput rate rather than under ant population size is the fact that many manufacturing systems which use closed loop input will attempt to produce at the rate at which products are anded and will choose population sizes accordingly.

ie results are tabulated for three different throughput rates which correspond 3.9%, 63% and 99.4% utilization levels. At low and medium utilization levels ie policies performed equally well but at the utilization level of 99.4% WBAL y outperforms all the other policies. Further, table 5 shows WBAL policy ves the desired through put at lower population sizes compared with the other ies under heavy loading conditions.

*Scheduling a two-station network with controllable inputs*

scheduling problem is relevant for any production system which is obliged to tain a specific average throughput rate of a certain product mix but can exert rol on the timing of inputs. Make-to-stock production systems stand as an ple to such situations. Advantage of controlling inputs lies in the fact that it lts in considerable reduction in WIP and in cycle times, there by improving the bility of the system.

ere we consider a simple case of a two station network with an endless queue of omers waiting to get entry into the system. Each customer has an exogenously ified class designations which are assigned such that the long-run proportion of s  $k$  customers released into the system is  $q_k$  for  $k = 1, \dots, K$ , satisfying,

$$\sum_{k=1}^K q_k = 1$$

input decision allows full discretion over timing of release of customers into system but no control can be exerted on the choice of which class to inject. thermore, there is a lower bound  $\bar{\lambda} = (\lambda_k)$ ,  $k = 1, \dots, K$  on the long-run

Table 5. Simulation results for the model in figure 2.

Throughput rate = .0274			
Policy	Mean Cycle Time	Var. of Cycle time	Population
WBAL	72.88	40.82	2
FBFS	72.88	40.82	2
FCFS	72.88	40.82	2
LBFS	72.88	40.82	2
SEPT	72.88	40.82	2
Throughput rate = 0.0433			
WBAL	138.57	206.99	6
FBFS	415.59	1789.51	18
FCFS	207.6	178.88	8
LBFS	161.58	93.31	7
SEPT	161.67	128.08	7

average throughput rate. Holding cost of  $c_k$  is incurred per unit time the class  $k$  spends in the system; but no set up costs are incurred during switchovers.

Let  $\{N(t), t \geq 0\}$  denote the input process which at any time  $t$  gives the cumulative number of customers released into the network during the interval  $[0, t]$ . Scaling and centering of the input process yields  $\theta(t)$  as,

$$\theta(t) = n^{-1/2} [\bar{\lambda} n t - N(nt)].$$

Following the notation of §3.1, the Brownian network formulation for the problem described is as follows:

(P.4.3.1):

Choose a pair  $(Y, \theta)$  to

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T \sum_{k=1}^K c_k Z_k(t) dt \right]$$

subject to

$$Y \text{ and } \theta \text{ are nonanticipating w.r.t } X \quad (148)$$

$$Z(t) = X(t) + RY(t) - q\theta(t), \quad \forall t \geq 0 \quad (149)$$

$$U(t) = AY(t), \quad \forall t \geq 0 \quad (150)$$

$$U \text{ is nondecreasing with } U(0) = 0 \quad (151)$$

$$Z(t) \geq 0, \quad \forall t \geq 0 \quad (152)$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [U_i(t)] \leq \gamma_i, \quad i = 1, 2. \quad (153)$$

where

$$\gamma_i = \sqrt{n} (1 - \rho_i). \quad (154)$$

Constraint (153) is a surrogate constraint to stipulate that the longrun average throughput should be greater than or equal to  $\bar{\lambda}$ . As in the open network case (see §3.1), because of non-singularity of the input-output matrix  $R$ , there exists a unique non-negative  $K$ -dimensional vector  $\beta = (\beta_k)$  satisfying flow balance equations,

$\lambda \equiv q \bar{\lambda}$ .

the vector of traffic intensities  $\rho = (\rho_i)$ ,  $i = 1, 2$ , and the vector of workload portions  $\alpha = (\alpha_k)$ ,  $k = 1, \dots, K$  are defined as in §3.1. The drift  $\delta$  and covariance matrix  $\Sigma$  of the Brownian motion  $X$  can be computed as discussed in §3.1.

The workload profile matrix  $M$  is defined by,

$$M = A R^{-1}$$

gives the expected total amount of time that server  $i$  must devote to a class  $k$  customer before it exits from the system. However, to find out total workload in the system, variations due to input control need be accounted for.

To this end, define the two dimensional vector  $v = (v_i)$  by

$$v = M q$$

so that  $v_i$  can be interpreted as expected total amount of time the server  $i$  devotes to each customer.

Now, the two dimensional scaled workload process defined by,  $W(t) = M Z(t)$  plus the additional workload due to input control, given by,  $v \theta(t)$ , gives the total workload in the system at any time  $t$ . Thus, the workload formulation for (P.4.3.1)

is:

use RCLL processes  $(Z, U, \theta)$  to

$$\text{minimize} \quad \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T \sum_{k=1}^K c_k Z_k(t) dt \right]$$

subject to

$$Z, U, \text{ and } \theta \quad \text{are nonanticipating w.r.t } X \quad (155)$$

$$U \quad \text{is nondecreasing with } U(0) = 0 \quad (156)$$

$$Z(t) \geq 0, \quad \forall t \geq 0 \quad (157)$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [U_i(T)] \leq \gamma_i, \quad \text{for } i = 1, 2 \quad (158)$$

$$M Z(t) + v \theta(t) = B(t) + U(t), \quad \forall t \geq 0 \quad (159)$$

where  $B(t)$  is defined by,

$$B(t) = M X(t)$$

and thus, has drift  $M\delta$  and covariance matrix  $M\Sigma M^T$ . For a proof of the equivalence of (P.4.3.1) and (P.4.3.2), refer Wein (1990b).

Given the policy  $U(t)$  at any fixed time  $t$ , embedded in (P.4.3.2) is a linear programming problem in  $Z$  and  $\theta$ . As the RHS of the constraint set of (P.4.3.2) varies with  $t$ , it would be easier to consider the corresponding dual LP, which has a static constraint set. Thus, the dual program for (P.4.3.2) is:

$$\text{maximize}_{\pi_1(t), \pi_2(t)} [B_1(t) + U_1(t)] \pi_1(t) + [B_2(t) + U_2(t)] \pi_2(t)]$$

subject to

$$M_{1k} \pi_1(t) + M_{2k} \pi_2(t) \leq c_k, \quad \forall k = 1, \dots, K \quad (160)$$

$$v_1 \pi_1(t) + v_2 \pi_2(t) = 0. \quad (161)$$

It can be shown easily that  $\rho_i = v_i \bar{\lambda}$ ,  $i = 1, 2$ . This fact renders it possible to simplify the dual LP further. To see this, define the workload imbalance process  $\widehat{W}(t)$  by,

$$\widehat{W}(t) = \rho_2 W_1(t) - \rho_1 W_2(t), \quad \forall t \geq 0 \quad (162)$$

Then, the dual LP reduces to,

$$\text{maximize}_{\pi_1(t)} \left[ \frac{\widehat{W}(t)}{\rho_2} \pi_1(t) \right]$$

subject to

$$c_k^{-1} (\rho_2 M_{1k} - \rho_1 M_{2k}) \pi_1(t) \leq \rho_2. \quad (163)$$

Order the classes  $k = 1, \dots, K$  so that

$$\arg_k \max c_k^{-1} (\rho_2 M_{1k} - \rho_1 M_{2k}) = 1 \quad (164)$$

$$\arg_k \min c_k^{-1} (\rho_2 M_{1k} - \rho_1 M_{2k}) = 2. \quad (165)$$

From the complementary slackness condition, it follows that,

$$Z_k(t) = 0, \quad \forall k \neq 1, \text{ if } \widehat{W}(t) > 0 \quad (166)$$

$$Z_k(t) = 0, \quad \forall k \neq 2, \text{ if } \widehat{W}(t) < 0. \quad (167)$$

Using this, it is easy to derive that when  $\widehat{W}(t) > 0$ ,

$$Z_k(t) = \begin{cases} \frac{\widehat{W}(t)}{\rho_2 M_{11} - \rho_1 M_{21}} & \text{if } k = 1 \\ 0 & \text{if } k \neq 1, \end{cases}$$

$$Z_k(t) = \begin{cases} \frac{\widehat{W}(t)}{\rho_2 M_{12} - \rho_1 M_{22}} & \text{if } k = 2 \\ 0 & \text{if } k \neq 2. \end{cases}$$

Thus, the optimal queue length process  $Z(t)$  does not depend on the control process  $\theta$  and depends on the control process  $U$  only through the workload imbalance process. The cost function corresponding to the optimal queue length process is given as a function of  $\widehat{W}(t)$ , i.e.,

$$\sum_{k=1}^K c_k Z_k(t) = h(\widehat{W}(t))$$

where,

$$h(x) = \begin{cases} -h_1 x & \text{if } x < 0 \\ h_2 x & \text{if } x > 0. \end{cases}$$

with  $h_1 = c_2/(\rho_1 M_{22} - \rho_2 M_{12})$  and  $h_2 = c_1/(\rho_1 M_{11} - \rho_2 M_{21})$ .

Hence, the workload problem is reduced to finding out optimal two dimensional cumulative idleness process,  $U$ . Further simplification is possible if we define the Brownian motion  $\widehat{W}$  and the right and the left control processes  $R$  and  $L$  by,

$$\widehat{W} \equiv \rho_2 B_1(t) - \rho_1 B_2(t), \quad \forall t \geq 0 \quad (168)$$

$$R(t) \equiv \rho_2 U_1(t), \quad \forall t \geq 0 \quad (169)$$

$$L(t) \equiv \rho_1 U_2(t), \quad \forall t \geq 0 \quad (170)$$

Then,  $\widehat{W} = \widehat{B}(t) + R(t) - L(t)$ ,  $\forall t \geq 0$ . Further, notice that  $\widehat{W}$  has drift  $\mu = \sqrt{n}(\rho_1 - \rho_2)$ . Using the relation (154), it is easy to see that,

$$\gamma_i = \frac{(1 - \rho_i) \mu}{\rho_1 - \rho_2}.$$

Now, the limiting control problem is,

(P.4.3.4):

Choose a pair  $(R, L)$  to

$$\text{minimize } \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T h(\widehat{W}(t)) dt \right]$$

subject to

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [R(t)] \leq \frac{\rho_2 (1 - \rho_1) \mu}{\rho_1 - \rho_2} \quad (171)$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [L(t)] \leq \frac{\rho_1 (1 - \rho_2) \mu}{\rho_1 - \rho_2}. \quad (172)$$

The problem (P.4.3.4) can be solved using *Lagrangian Multipliers* method. For this we need the following *Lagrangian cost function*:

$$K(x) = \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T h(\widehat{W}(t)) dt + r R(T) + l L(T) \right] \quad (173)$$

where  $r$  and  $l$  are the Lagrangian multipliers corresponding to the constraints (171) and (172). Call this problem as *Lagrangian problem*. With the aid of the following theorem, the constrained problem (P.4.3.4) can be solved by making an appropriate choice of multipliers and then minimizing the cost function  $K(x)$ .

**Theorem 4.3.1** Suppose  $r$  and  $l$  are nonnegative real numbers and suppose  $(R^*, L^*)$  is a solution to the Lagrangian problem. Furthermore, suppose

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [R^*(t)] = \frac{\rho_2 (1 - \rho_1) \mu}{\rho_1 - \rho_2} \quad (174)$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [L^*(t)] = \frac{\rho_1 (1 - \rho_2) \mu}{\rho_1 - \rho_2}. \quad (175)$$

Then,  $(R^*, L^*)$  is a solution to the constrained control problem (P.4.3.4).

See Wein (1990a). Taksar (1985) developed sufficient conditions for the optimality of the Lagrangian problem. The optimal policy is one among a special class of policies called *control limit policies*. Such a policy brings the controlled process  $\widehat{W}(t)$  within a certain interval  $[a, b]$  instantaneously and keeps it within that interval while exerting minimum amount of control. The process  $\widehat{W}(t)$  under such a policy is an RBM in the interval  $[a, b]$ .

The control limit policy on  $[a, b]$  is defined by,

$$R(t) = \sup_{0 \leq s \leq t} [a - \widehat{B}(t) + L(s)]^+ \quad (176)$$

$$L(t) = \sup_{0 \leq s \leq t} [\widehat{B}(s) + L(s) - b]^+. \quad (177)$$

Taksar (1985) gives sufficiency conditions for a control limit policy on  $[a, b]$  to be a solution to the Lagrangian problem. Wein (1990a) using this result and the theorem (4.3.2) derives sufficiency conditions for a control limit policy to be a solution to problem (P.4.3.4). Thus, the problem is reduced to finding out candidate interval end points  $a^*$  and  $b^*$  corresponding to  $R^*$  and  $L^*$  of the theorem (4.3.2).

In order to find these points  $a^*$  and  $b^*$ , the following lemma from Harrison (1985) is needed.

*Lemma 4.3.1 Let  $\widehat{B}$  be a  $(\mu, \sigma^2)$  Brownian motion and  $R$  and  $L$  be as in (176) and (177) and thus,  $\widehat{W} = B + R - L$  is an RBM on the interval  $[a, b]$ . Then,  $\widehat{W}$  has truncated exponential steady state distribution with density,*

$$p(x) = \frac{\nu e^{\nu(x-a)}}{e^{\nu(b-a)} - 1} \quad \text{for } a \leq x \leq b. \quad (178)$$

where  $\nu = \frac{2\mu}{\sigma^2}$ . Furthermore,

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [R(T)] = \frac{\mu}{e^{\nu(b-a)} - 1} \quad (179)$$

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [L(T)] = \frac{\mu}{1 - e^{\nu(b-a)}} \quad (180)$$

In view of the theorem (4.3.1), the interval end points can be found by solving the following problem:

(P.4.3.5):

Among the class of control limit policies, find a policy  $(R, L)$  to

$$\text{minimize} \quad \limsup_{T \rightarrow \infty} \frac{1}{T} E_x \left[ \int_0^T h(\widehat{W}(t)) dt \right]$$

subject to,

$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [R(T)] = \frac{\rho_2(1 - \rho_1)\mu}{\rho_1 - \rho_2} \quad (181)$$

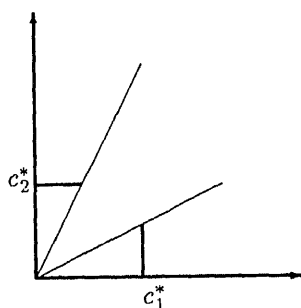
$$\limsup_{T \rightarrow \infty} \frac{1}{T} E_x [L(T)] = \frac{\rho_1(1 - \rho_2)\mu}{\rho_1 - \rho_2} \quad (182)$$

The above lemma enables to express the constraints in (P.4.3.5) directly in terms of the end points  $a$  and  $b$  and thus, establishes a relation between  $a$  and  $b$ . As a result, the problem reduces to a search over values of  $a$ .

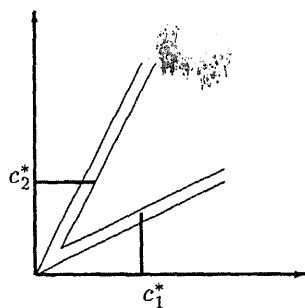
For a detailed description of the solution procedure, refer Wein (1990a). Now, in the optimal solution only class 1 customers have positive queue length whenever  $\widehat{W}(t) > 0$  and class 2 customers have positive queue length whenever  $\widehat{W} < 0$ . This can be interpreted to mean that customers of class 1 are given the lowest priority whenever  $\widehat{W}(t) > 0$ . As for the priority of the other classes, a natural policy would be to award highest priority at each station at any time  $t$  to the customer with largest reduced cost. The reduced cost for a class  $k$  customer at time  $t$  gives the increase in the objective function value of the problem (P.4.3.3) per unit increase in RHS of the corresponding constraint. These dynamic reduced costs,  $c_k$  ( $k = 1, \dots, K$ ) can be easily found out from the dual program (P.4.3.4). For further details, see Wein

90b). Since  $w = MZ$ , the optimal solution  $Z^*$  implies that the workload process resides on the boundary of a cone in  $R^2_+$ . Further, the optimal control policies  $R^*$  and  $L^*$  are such that the control,  $U_1$  (respectively,  $U_2$ ) is exerted only when  $\widehat{W}(t) = a^*$  (respectively,  $b^*$ ). In other words, idleness is incurred when  $\widehat{W} = a^*$ . This can be expressed in terms of the workload process  $W(t)$  using the optimal queue length process.

The interval end points  $a^*$  and  $b^*$  are the reflecting barriers on the boundary of the cone beyond which  $W(t)$  may not enter.  $W(t)$  must reside on a portion of the boundary of the cone as shown in figure 3. The optimal solution tells that controls  $U_1$  and  $U_2$  are exerted only when  $W_1(t) = c_1^*$  and  $W_2(t) = c_2^*$  respectively, (See figure 3). Otherwise, only the input process  $\theta(t)$  is used to keep the workload process on the boundary of the cone. To be more precise, input is increased relative to the nominal input rate whenever the process  $W(t)$  lies in the shaded regions and input is withheld whenever  $W(t)$  is inside the cone.



**Figure 3.** Cone of confinement for the process  $W(t)$ .



**Figure 4.** Inner cone to adopt input control.

However, in the actual queueing system, the process  $W(t)$  may reside outside the cone in figure 3. This is because the state space of the workload process is the cone  $W = MZ$ ,  $Z \geq 0$  and its extremal rays are generated by the two customer classes with

$$\arg \max_k \frac{M_{1k}}{M_{2k}}$$

$$\arg \min_k \frac{M_{1k}}{M_{2k}}$$

which may not coincide with the classes 1 and 2 of the priority rule described earlier.

In the idealized Brownian model, when the scaled workload process is on the lower ray and  $W_2(t) < c_1^*$ , then there are zero customers at station 1; but station 1 is not idle according to the input rule described above. This apparent paradox is due to the scaling process involved in heavy traffic limit. Eventhough, in the actual system there are enough customers at station 1, these customers vanish in the scaled space of heavy traffic.

Table 6. Data for the model in figure 5.

Class	Service Dist.
1	Uniform(2.0, 1.7)
2	Normal(5.0, 1.0)
3	Exp(4.0)
4	Exp(8.0)
5	Normal(6.0, 1.0)
6	Exp(9.0)

In order to adopt the input rule to the actual system, it is necessary to consider a cone which is generated from the original one by building up a boundary layer of thickness, say  $\epsilon$ , (see figure 4), inside the original cone. Now, the input rule admits customers as long as the workload process is in the enlarged shaded area. Selection of such a suitable  $\epsilon$  is dependent on the network topology and also on how balanced the network is. Further, in the whole description given above, we have considered only the *scaled* workload process  $W(t)$ . In order to adopt the policy, this has to be reexpressed in unscaled terms. The procedure is the same as that has been done in the make-to-stock case. (See §4.1) and for further details, interested readers are referred to Wein (1990b). The case of multi-station closed networks is discussed in detail in Wein (1992a).

#### 4.3.1 An example: A two-station re-entrant line.

A two station re-entrant line shown in figure 5 is considered for the performance study of workload regulating release policy and dynamic reduced cost based priority sequencing policy through simulation. The service distributions for the classes shown in figure 5 are given in table 6. All  $c_i$ 's are assumed to be equal to 1.0. To achieve a throughput rate of 20 jobs per unit time, the values of  $c_1^*$  and  $c_2^*$  (see figure) should be 87.6 and 56.3 respectively. The boundary layer thicknesses  $\epsilon_1$  and  $\epsilon_2$  are set at 1.0. Different combinations of input release policies and priority sequencing policies are experimented. The results are presented in table 7.

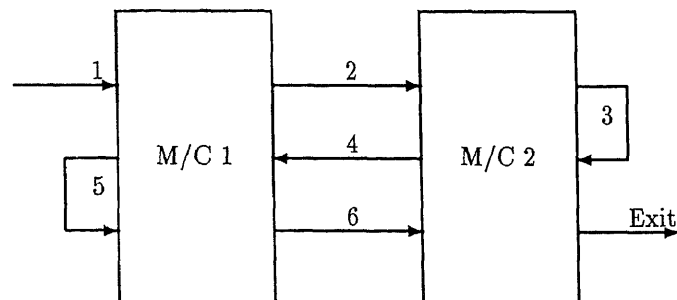


Figure 5. A two station re-entrant line



Table 7. Simulation results for the model in figure 5.

policy	Policy	Time	Cycle Time
DRC	WR	137.3	1499.4
FCFS	Deterministic	184.3	5861.1
SRPT	Deterministic	172.0	4135.2
LBFS	Deterministic	182.7	3842.6
FBFS	Deterministic	181.1	10397.0
FCFS	Poisson	255.0	16544.8
SRPT	Poisson	227.8	12025.9
LBFS	Poisson	228.9	12025.9
FBFS	Poisson	372.9	53683.6
FCFS	Uniform	190.7	7292.3
SRPT	Uniform	181.2	5398.4
LBFS	Uniform	188.7	4560.2
FBFS	Uniform	191.4	13591.3

One can see that DRC in association with WR release policy performed better than other policies in reducing cycle times where as under poisson input release policy, large cycle times are incurred. As poisson release policy can be thought of as representative to open loop release policy that is independent of the state of the system, one can say that by exercising control over input release, and thus regulating the amount of work at bottleneck stations, WIP levels can be reduced considerably.

## 5. Future work

The use of Brownian motion in the dynamic scheduling of multiclass queueing networks is now well established. In the past five years, there is a large body of literature on this subject. Much of the early work focussed on single station and two station networks (see Harrison & Wein 1989, 1990) and Wein (1990b, 1992b), but results are now available for multi-station networks (see Wein 1992a), Chevalier & Wein (1993). On the theoretical front, a heavy traffic limit theorem for very general networks is still eluding researchers. Also, recently, multiclass queueing networks that do not have a satisfactory Brownian network approximation in heavy traffic have been presented (Dai & Vien Nguyen 1992). This explains the need for characterizing classes of networks having Brownian networks that approximate them satisfactorily under heavy traffic assumptions. It is also an interesting open issue to investigate the range of values of traffic intensities for which a given network can be satisfactorily approximated by a Brownian network.

There are several interesting scheduling problems that can be attempted using Brownian approximations. There is a large variety of scheduling problems that one sees in the real world since every factory or manufacturing setup has its own peculiar and unique scheduling problems. The following gives a list of real-world features that are worthwhile to be taken into account while scheduling resources in a manufacturing facility.

- Delayed or stochastic availability of raw material: Since the raw materials are usually procured from sources external to the machine shop, one is never sure unless the raw material is in hand. Brownian models have so far assumed

a perennial supply of raw material and also do not account for raw material holding cost.

- In a real-life factory, machines or tools are prone to breakdowns and these events are non-deterministic. Also, the processed parts need not always conform to the required quality standards. Usually, periodic inspection of processed parts will decide whether the quality of parts is acceptable or not. Parts identified for reworking cause extra load on the system whereas every rejected part entails complete reprocessing and also material waste. Modelling such features is important.
- In a multiclass production system, switchover times or set-up times can have a significant effect on the way parts are scheduled. Existing Brownian models do not address the issue of scheduling in the presence set-up times and set-up costs.
- The objective function chosen for minimization in the existing literature usually takes into account factors such as inventory costs, backorder costs, mean waiting times, and machine utilizations. Since variability of performance measures is also a very important criterion, there is need to include it as part of the objective functions. There are also other measures of performance such as makespan and total tardiness. This also brings out the issue of modeling due dates.

It is also important and useful to evaluate the performance of Brownian policies and derive performance degradation of Brownian policies when the underlying network does not satisfy heavy traffic conditions. Finally, one has to look into the computational effort involved in arriving at Brownian policies for various networks.

This research was supported by the Office of Naval Research and the Department of Science and Technology grant N00014-93-1017. We would also like to acknowledge the encouragement and comments of Professor N. Viswanadham and several critical comments of the reviewers of this paper.

## References

- Abdul-Razaq T S, Potts C N 1988 Dynamic programming state space relaxation for single machine scheduling. *J. Oper. Res. Soc.* 39: 141-152
- Bagchi U, Ahmadi R H 1987 An improved lower bound for minimizing weighted completion times with deadlines. *Oper. Res.* 35: 311-313
- Baker K R, Ahmadi R H 1978 Finding an optimal sequence by dynamic programming: An extension to precedence-related tasks. *Oper. Res.* 35: 111-120
- Beloudah H, Posner M E, Potts C N 1988 A branch and bound algorithm for scheduling jobs with release dates on a single machine to minimize total weighted completion time. Preprint, Faculty of Mathematical Studies, University of Southampton

- Billingsley P 1968 *Convergence of probability measures* (New York: John Wiley and Sons)
- Breiman L 1968 *Probability* (Reading, MA: Addison-Wesley)
- Chevalier P B, Wein L M 1993 Scheduling network of queues: Heavy traffic analysis of a multistation closed network. *Oper. Res.* 41: 743-758
- Cox D R, Miller H D 1965 *The theory of stochastic processes* (London: Methuen)
- Dai J G, Vien Nguyen 1992 On the convergence of multiclass queueing networks in heavy traffic. School of Mathematics and Industrial/Systems Engineering, Georgia Institute of Technology
- Fisher M L 1973 Optimal solution of scheduling problems using Lagrangian multipliers. *Oper. Res.* 21: 1114-1127
- Fisher M L 1981 Lagrangian relaxation method for integer programming problems. *Manage. Sci.* 27: 1-18
- Flores C 1985 Diffusion approximations for computer communication networks. *Proc. Symp. Appl. Math.* : 83-124
- Forst F G 1984 A review of the static, stochastic, job sequencing literature. *Opsearch* 21: 127-144
- Gere W S 1987 Heuristics in job shop scheduling. *Manage. Sci.* 13: 167-190
- Goldberg D E 1986 *Genetic algorithms in search, optimization, and machine learning*. (Reading, MA: Addison-Wesley)
- Hajek B 1974 Optimal control of two interacting service stations. *IEEE Trans. Autom. Contr.* 29: 491-499
- Harrison J M 1985 *Brownian motion and stochastic flow systems* (New York: John Wiley and Sons)
- Harrison J M 1988 Brownian models of queueing networks with heterogeneous customer populations. *Stochastic differential systems, stochastic control theory and applications, IMA* (eds) W Fleming, P L Lions (New York: Springer-Verlag) 10: 147-186
- Harrison J M, Reiman M 1981 Reflected Brownian motion on an orthant. *J. Appl. Probab.* 9: 302-308
- Harrison J M, Wein L M 1989 Scheduling network of queues: Heavy traffic analysis of a simple open network. *Queueing Syst.: Theor. Appl.* 5: 265-280
- Harrison J M, Wein L M 1990 Scheduling network of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* 38: 1052-1064
- Harrison J M, Williams R J 1987 Multidimensional reflected Brownian motion having exponential stationary distributions. *Ann. Probab.* 15: 115-137

- Hochbaum D S, Shmoys D B 1989 A polynomial approximation scheme for machine scheduling on uniform processors using the dual approximation approach. *SIAM J. Comput.* 17: 539-555
- Karlin S, Taylor H M 1981 *A first course in stochastic processes* (New York: Academic Press)
- Klimov G P 1974 Time sharing service systems. *Theor. Appl. Probab. Appl.* 19: 532-551
- Lawler J K, Lenstra J K, Rinooy Kan A H G, Shmoys D B 1990 Sequencing and scheduling: Algorithms and complexity. *Handbook of operational research and management* (eds) S C Graves, A H G Rinooy Kan, P Zipkin (Amsterdam: North-Holland) 4: 51-61
- Lemoine A J 1978 Network of queues: A survey of weak convergence results. *Management. Sci.* 24: 175-1193
- Levy B C, Adams M B 1987 Global optimization with stochastic neural networks. *Proceedings of IEEE International Conference on Neural Networks*, San Diego, pp 681-686
- Pinedo M L 1981 A note on the two-machine job shop with exponential processing times. *Naval Res. Logist. Q.* 28: 693-696
- Pinedo M L 1982 Minimizing the expected make-span in stochastic flow shops. *Oper. Res.* 30: 148-162
- Pinedo M L 1983 Stochastic scheduling with release dates and due dates. *Oper. Res.* 31: 559-572
- Pinedo M L, Scrage L 1982 Stochastic shop scheduling: A survey. *Deterministic and stochastic scheduling* (eds) M A H Dempster, J K Lenstra, A H G Rinooy Kan (Dordrecht: Reidel) pp 181-196
- Pinedo M L, Weiss G 1980 Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. *J. Appl. Probab.* 17: 187-202
- Pinedo M L, Weiss G 1987 The "Largest variation first" policy in some stochastic scheduling problems. *Oper. Res.* 35: 884-894
- Pollard D 1984 *Convergence of stochastic processes* (New York: Springer-Verlag)
- Reiman M I 1982 Heavy traffic diffusion approximation for sojourn times in Jackson networks. *Applied probability, computer science, the interface* (eds) R L Disney, T T Ott (Boston: Birkhauser) pp 409-422
- Reiman M I 1984 Open queueing networks in heavy traffic. *Math. Oper. Res.* 9: 441-458
- Ross K W, Yao D D 1989 Optical dynamic scheduling in Jackson networks. *IEEE Trans. Autom. Contr.* 34: 47-53

- Stidham S Jr 1985 Optimal control of admission to a queueing system. *IEEE Trans. Autom. Contr.* 30: 44-52
- Taksar W 1985 Average optimal singular control and a related stopping problem. *Math. Oper. Res.* 10: 63-81
- Van Laarhoven P J M, Aarts E H L, Lenstra J K 1992 Job shop scheduling by simulated annealing. *Oper. Res.* 40: 1156-1179
- Veatch M H, Wein L M 1992 Scheduling a make-to-stock queue: Index policies and hedging points. Working Paper, Massachusetts Institute of Technology, September
- Walrand J 1988 *An introduction to queueing networks* (Englewood Cliffs, NJ: Prentice Hall)
- Weber R R, Stidham S 1987 Optimal control of service rates in network of queues. *Adv. Appl. Probab.* 19: 202-218
- Wein L M 1990a Optimal control of a two-station Brownian network. *Math. Oper. Res.* 38: 215-242
- Wein L M 1990b Scheduling network of queues: Heavy traffic analysis of a two-station network with controllable inputs. *Oper. Res.* 38: 1065-1078
- Wein L M 1992 Scheduling network of queues: Heavy traffic analysis of a multi-station network with controllable inputs. *Oper. Res.* 40: S312-S334
- Weiss G 1982 Multiserver stochastic scheduling. *Deterministic and stochastic scheduling* (eds) M A H Dempster, J K Lawler, A H G Rinnooy Kan (Dordrecht: Reidel) pp 157-179
- Whitt W 1974 Heavy traffic limit theorems for queues: A survey. *Mathematical methods in queueing theory* (ed.) A B Clarke (New York: Springer-Verlag) pp 307-350
- Whitt W 1985 Some useful functions for functional central limit theorem. *Math. Oper. Res.* 5: 67-85
- Whitt W 1982 Refining diffusion approximations for queues. *Oper. Res. Lett.* 1: 165-169
- Wolff R A 1989 *Stochastic processes and queueing theory* (Englewood Cliffs, NJ: Prentice Hall)
- Yao D D, Shantikumar J G 1990 Optimal scheduling control of a flexible machine. *IEEE Trans. Robotics Autom.* 6: 706-712



# Design of software for safety critical systems

R K SHYAMASUNDAR

Tata Institute of Fundamental Research, Bombay 400 005,  
India

**Abstract.** In this paper, we provide an overview of the use of formal methods in the development of safety critical systems and the notion of *safety* in the context. Our attempt would be to draw lessons from the various research efforts that have gone in towards the development of robust/reliable software for safety-critical systems. In the context of India leaping into hi-tech areas, we argue for the need of a thrust in the development of quality software and also discuss the steps to be initiated towards such a goal.

**Keywords.** Formal methods; safety-critical systems; software design; synchronous programming paradigm.

*"If only we could learn the right lessons from the successes of the past,  
we would not need to learn from our failures"*

C.A.R. Hoare

## 1. Introduction

Historically and traditionally (Simon 1969) it has been the task of the science disciplines to teach about natural things: *how they are and how they work*. It has been the task of engineering schools to teach about artificial things: *how to make artifacts that have desired properties and how to design*. Webster's dictionary defines engineering as "the application of scientific principles to practical ends as the design, construction, and operation of efficient and economic structures, equipment and systems". By this definition, Computer Science can be viewed as engineering with the "design of programs" as one of the principle activities. Like many other professions, *design* happens to be the core of the engineering profession. It is surprising that one does find a discipline that could be called "philosophy of design" as a counterpart to "philosophy of science" – a well established discipline traditionally. As Herbert Simon argues, that the emergence of the activity of "Design of Programs" in computer science (usually termed *Software Engineering*) has also paved the way for "The Sciences of Design". The major aim of software engineering is to

---

\*An earlier version was presented as an Invited paper at the ISRO Conference on Software Engineering, VSSC, Trivandrum, 29-30 July 1994.

direct the enormous resources of computational power on the silicon chip to the use and convenience of mankind.

Metaphor and analogy can be helpful, or they can be misleading. All depends on whether the similarities the metaphor captures are significant or superficial and ignore the underlying reality. As it stands, significant amount of research criteria for Computer Science has been borrowed from adjacent disciplines of Science, particularly Mathematics. For instance, in mathematics novelty and consistency are the main criteria for measurement of relevance or success rather than applicability<sup>1</sup>. The analogy between programming and traditional engineering disciplines has been very fruitful, and has provided the much needed basis and advantages. However, there are fundamental differences between software and other technologies. Some of the major differences arise due to distinct notions of complexity measures, analysis of reliability, usage of tools, standards etc. in software and other technologies (Hoare & Jones 1989; Parnas *et al* 1990). In that sense, analogy between software and other engineering disciplines breakdown on the following fronts (Hoare & Jones 1989; Parnas *et al* 1990).

1. Complexity measure: Software and hardware differ in the measures of complexity, be it design, development or usage.
2. Methods of achieving reliability: It appears that we should not have any difficulty in achieving reliability while designing software. The reasons are that basic raw materials for programs (registers, bytes, disks, tapes etc.) are almost unbounded and programs work in a controlled environment; there is no need to worry about defective components, friction, unskilled labourers, natural catastrophes like storms, earthquakes etc. This is not the case for the following two main reasons:
  - (a) The most important way of achieving reliability in a product is *testing* under extreme conditions (perhaps taking into account a factor of safety) such as temperatures, pressure, voltage (essentially continuous variables) etc. However, in the case of software testing there is no analogous procedure of testing; establishing that the program works for the boundary values is no guarantee that it works for values in between. In some sense, the methods of extrapolation and interpolation that come to rescue in the traditional testing does not help at all as far as program designs are concerned. Lack of *continuity of behaviour with respect to input* in the sense of the traditional engineering products adds an additional dimension to the problems.
  - (b) The discipline of software design is still immature and anyone who can type on a keyboard appears to have the impression that he/she can program. This has led to an attitude that there is no need to pay due atten-

<sup>1</sup>One may recall a quotation from Christopher Strachey in connection with the setting of school at Oxford in 1974: "It has long been my personal view that the separation of practical and theoretical work is artificial and injurious. Much of the practical work done in computing, both in software and in hardware design, is unsound and clumsy because the people who do it do not have any clear understanding of the fundamental principles underlying their work. Most of the abstract mathematics and theoretical work is sterile because it has no point of contact with real computing. One of the central aims of the Programming Research Group, as a teaching and research group, has been to set up an atmosphere in which this separation cannot happen ..."



tion to the underlying mathematical abstractions. Design is developed on the fly, without much careful analysis and review.

3. Modular structure: The notion of *modules* in engineering is more often used in the sense of *independent* units spatially separated; that is, modules are not likely to interfere in the functioning of other modules. However, the situation differs when it comes to programming as there are no such spatial separations. In fact, even assuming the best *modular* approach, such a separation may not be possible due to efficiency considerations leading to practices such as code sharing techniques.
4. Tools: A programmer has to use the tools such as compilers (for programming languages), editors, environments etc. on which he does not even have a clear understanding (perhaps, they may even have bugs!). Above all he has to deal with software manuals which are more often than not highly unsatisfactory.
5. Evaluation: In spite of the fact that a large number of design methods have been proposed and a large number of systems has been built and used, one hardly finds evaluations that carefully demonstrate the success/failure of a system or of a method used in designing a system.

This paper is an attempt to draw lessons from the various research efforts that have gone into the development of reliable software for safety-critical systems.

The paper is organized as follows: Sections 2 and 3 provides overviews of formal methods and reactive/real-time systems respectively; this is followed by a systematic method of designing reactive systems in §4. Section 5 discusses case studies in the development of safety critical systems using formal methods followed by a discussion on the lessons drawn from the above experience in §6. Section 7 discusses one of the successful paradigms for reactive programming, namely the *synchronous paradigm*. In §8 we discuss the notion of *safety* in the design of safety-critical systems. In the concluding section, we discuss the steps needed to be taken in India for the development of reliable reactive systems in particular and reliable software in general.

## 2. Formal methods of software development

A method is said to be *formal* if it has a sound mathematical basis provided by its specification languages. Its main function is to check the consistency/completeness of the designers intentions and check whether it is realizable/implementable. It also provides a means of verifying whether the implementation meets its requirements and establish properties of the system without actually running the system. Of course, it is important that the method addresses the pragmatics of the designers; in fact, the success of any method also depends on how successfully the method addresses the various pragmatic considerations. Quite often, it is argued that the use of formal (mathematical!) methods is mandatory for improving the quality of software. The starting point for any formal method is the need for specification. Some of the important reasons for the need of specification are:

1. It serves as a contract (Lamport 1983) between the user and the implementor of the module. This serves to clearly articulate separate the following concerns:

- (a) Implementor: The implementor's responsibility is to meet the requirements imposed by the specification. Thus, there is no need for the implementor to know how the rest of the system works.
- (b) User: The user can use the module as a black box either in the use or the development of programs using this module.
- (c) Test: It precisely brings out testing criteria of the implementation.

2. As a communication among the implementors of the system.

3. Support the development of multi-version software.

Most of the problems that arise either during the development stage or during the use of a program can be attributed to inaccuracies, ambiguities and incompleteness of the problem/solution. The use of mathematical (formal) methods of specification guards against ambiguities and inaccuracies and a properly chosen method also enables to overcome inconsistencies and incompleteness issues.

There have been a plethora of methods for the development of software. Principal criteria used for the classification of software development methods have been (Place *et al*1990):

1. Representation: The foremost task is to represent the designer's intent. Various representations are feasible based on:
  - (a) State-based/event-based specifications.
  - (b) Style of specification such as declarative or model-based.
  - (c) Abstraction features such as concurrency, nondeterminism.
  - (d) Handling properties such as safety or liveness properties.
2. Transformation: Having represented the intentions at some level, the next task is: how do we transform the specification into another one that is more detailed than the one we have already, preserving the correctness? This is a very crucial step which also reflects largely on safety or the reliability of the method. The questions one asks are:
  - (a) Does it support compositionality?
  - (b) Does it support *rigorous*<sup>2</sup> derivation?
  - (c) Is the refinement (or reification) completely mechanical?
  - (d) Is there a pragmatic interface of the tools with the users?
3. Validation: This part corresponds to raising the level of confidence of the system relative to an environment using various testing, simulation and verification techniques/tools. With reference to verification, questions such as consistency, completeness, equivalence, safety and liveness become pertinent.

The methods can broadly be categorized based on the class of the underlying specification languages. Specification languages can be broadly classified as:

<sup>2</sup>This is used to mean that it is not necessary to perform proof; however, if needed it could be performed.

1. Model oriented (Concrete types),
2. Property oriented (Abstract types), and
3. Type oriented (dependent types).

VDM, Z, Raise are typical examples of model-oriented languages; LARCH, ASL, CLEAR, OBJ EML etc. are typical examples of the class of property-oriented systems and Calculus of Constructions is a typical variety in the type-oriented category. Further categorization is possible based on the support of programming languages/styles and paradigms such as object-oriented, concurrency, functional etc. We will not go into details on the specification languages further in this paper.

VDM and Z have been in industrial use for quite some time. VDM originated from design methods to develop concrete data structures from abstract specifications and hence, supports directly development. On the other hand, Z supports development of *requirement specifications* (discussed in the sequel) and thus, has properties such as conjunction, disjunction and also negation for describing constraints. Schemas in Z provide a pragmatic support for refining specifications. These formalisms have been used for specifying large software systems. However, these formalisms do not directly support features such as concurrency, communication features and also the specification of liveness properties and time. Temporal logic based systems are property-oriented systems and have the power to specify the behaviour of systems in a declarative way; consistency check can be achieved by building models which corresponds to building finite state machines. It is here, many of the restrictions appear. Even here, concurrency does not appear directly and "real-time" could be added to arrive at various classes of real-time temporal logics. One of the difficulties with these formalisms is the state explosion problem that one encounters when dealing with finite-state formalisms. These formalisms have been widely used for specifying requirement specifications. The lack of concurrency and the constraints one needs to specify in the context of reactive systems has also lead to look for formalisms, particularly algebraic so that one can do simulation and also arrive at verification tools based on various notions of bisimulation. We discuss the latter approach for the development of reactive systems in the sequel.

### 3. An overview: Real-time reactive systems

Computers are being increasingly used in a variety of applications ranging from home appliances or laboratory instruments to process control systems, flexible manufacturing, flight control and tactical control in military applications. In fact, their use has become essential due to stringent service requirements and availability of inexpensive hardware. For example, flexible manufacturing is a special kind of real-time application where the behaviour of each manufacturing machine can be adapted *instantaneously* to continuously changing working conditions while still satisfying a global optimality criterion. In flight control systems, *real-time* automatic maneuvering is used for significant reduction of fuel consumption and also for tactical control over the target. Needless to say that safety and reliability are extremely important for such systems since a failure may result in economic, human and ecological catastrophes. The term *Embedded systems* (has been almost synonymous with real-time systems) have become popular through Ada<sup>3</sup>. The term *Embedded* in Embedded

<sup>3</sup>Ada is a trademark of the US Department of Defense (AJPO).

systems refers to the fact that these systems are embedded in larger systems whose primary purposes are not general computations; the main purpose is to provide the needed support to achieve the overall objective of the system. One of the common concepts that counter a majority of the process control *embedded* systems is that of providing continual feedback to an unintelligent environment. The continual demands of an unintelligent environment cause these systems to have relatively rigid and urgent performance requirements, such as real-time response requirements and *fail-safe* reliability requirements. It seems that this emphasis on performance requirements is what really characterizes embedded systems, and causes us to be more aware of their roles in their environments than we are for other types of systems. Table 1 provides an informal classification of systems, based on properties that show up at the requirements level.

**Table 1.** An informal classification of systems.

Type	Characteristics	Examples
Embedded systems	<ul style="list-style-type: none"> <li>• Special purpose (application)</li> <li>• Absolute performance requirements</li> </ul>	<ul style="list-style-type: none"> <li>• Industrial process control systems</li> <li>• Flight guidance systems</li> </ul>
Data-processing systems	<ul style="list-style-type: none"> <li>• Special purpose (application)</li> <li>• Relative performance requirements</li> </ul>	<ul style="list-style-type: none"> <li>• Batch business programs</li> <li>• Online data base systems</li> </ul>
Support systems	<ul style="list-style-type: none"> <li>• General-purpose</li> <li>• Relative performance requirements</li> </ul>	<ul style="list-style-type: none"> <li>• Operating systems</li> <li>• Software development tools</li> </ul>

Systems such as airline-reservation systems should probably be viewed as a combination of these types. In addition to the performance requirements, which have been already established as a major distinguishing factor, embedded systems are especially likely to have stringent resource requirements. These are requirements on the resources (mainly physical in this case) from which the system is constructed. This is because embedded systems are often installed in places (such as satellites) where weight, volume, or power consumption must be limited, or where temperature, humidity, pressure and other factors cannot be as carefully controlled as in the traditional machine room.

The interface between an embedded system and its environment tends to be complex, asynchronous, highly parallel and distributed. This is another direct result of the *process control* concept, because the environment is likely to consist of a number of objects which interact with the system and each other asynchronously in a parallel fashion. Furthermore, it is probably the complexity of the environment that necessitates computer support in the first place (consider an air-traffic-control system!). This characteristic makes the requirements difficult to specify in a way that is both precise and comprehensible.

Finally, embedded systems can be extraordinarily hard to test. The complexity of the system/environment interface is one obstacle, and the fact that these programs

often cannot be tested in their operational environments is another. It is not feasible to test flight-guidance software by flying with it, nor to test ballistic-missile-defense software under battle conditions.

Embedded systems are often used in critical applications where real-time/reactive response is essential. The main characteristics of the embedded systems are summarized below:

1. The primary purpose is to provide the needed support to achieve the overall objective of the system rather than general computations.
2. The system tends to be large, complex and can be extraordinarily hard to test.
3. The environment that the system interacts with is nondeterministic. That is, most of the times, there is no way to anticipate in advance the precise order of external events.
4. High speed external events (perhaps in parallel), must be able to affect the flow of control in the system easily.
5. The real-time behaviour must be controllable, e.g. the requests must be handled within certain time limits.
6. The system is a coordinated set of asynchronous distributed units.
7. The mission time is long. The system not only must deal with ordinary situations but also must be able to recover from some extraordinary ones.

#### 4. Design and development of reactive systems

The process of design may be viewed as an iterated transformation of a conceptual abstract functional description through refined levels of description till the emergence of an implementation. Between the concept and implementation (realization of the concept in hardware or software given the physical resources), there can be various stages such as Service requirements, Functional requirements, Architectural requirements, Performance requirements, Detailed Design etc. In fact, there is no agreement about the precise meaning of a stage and also it is not really clear as to where one stage ends and the next begins. However, informally one understands the various ranges from concept to implementation, and we refer to each stage as a *specification*. Assuming, one has a *good view of customer's/designer's intents*, the following two stages (Pnueli 1986) generally play a vital role in the development of systems:

1. Requirement specifications: This is perhaps the closest to the conceptual understanding of the problem (and thus, the earliest stage). In a sense, this could be viewed as the basis for a contract between a customer who orders the system, and a representative of an implementation team who is supposed to construct the system. The important parts of a requirement specifications are the *static part*, which identifies the interface between the system and its environment, and the *dynamic part*, which specifies the behaviour that the customer expects to observe on the identified interface. It may be noted that the identification of the interface depends on various architectural issues. In

short, the requirement specification concentrates on the observable behaviour between the customer and the various components of the system without the knowledge of the implementation (and hence, the internal structure) of the components of the system.

2. System specifications: This has to achieve the "how" part from the "what" part and the architectural considerations. It has basically the following concerns:

- Architecture of the system: The major concern here is to decompose the system into a set of (hierarchical!) subtasks.
- Mapping the logical architecture into the physical architecture.
- Interactions among the sub-tasks.

In a broad sense, system specifications are almost executable. However, it may be noted that a system specification need not represent an actual implementation.

If both a *requirement (service) specification* and an *implementation specification* have been constructed for a system, it is possible to validate the implementation specification by confirming that it satisfies the service specification. This ability is very valuable as the implementation specification is often quite complex and prone to error, while the service specification is much shorter and simpler.

#### 4.1 What's the concern of requirement specification?

By specifying the minimum required externally visible behaviour, and leaving all other aspects to lower levels of description, one can obtain a more general specification that reflects the necessary requirements of the system. A specification that is oriented towards one implementation may discourage or even preclude other equally valid implementations. Thus, this stage can be viewed as capturing either *the behaviour of the customer and the system as seen by an observer who can see the components (customer included) as a black box* or *the properties the system should satisfy*.

In a real-world, it must be evident that when a system is being developed it is generally the case that the requirements would have to be augmented either because an incompleteness was detected in the informal description of the customer/designer or the mistake/unsatisfactory performance was detected after the system was designed. Thus, the two most important concerns in this stage are

1. Consistency: A spectrum of examples can be seen where requirements contradict each other (it could be in the safety requirements or between the safety and eventual requirements). The *consistency* ensures that there is some execution model satisfying the requirements.
2. Completeness: The main concern here is: Does the stipulated requirements ensure *what the customer wants* or the requirements have to be strengthened to avoid *undesirable behaviour*?

Thus, logic-based formalisms would be appropriate for this stage. There have been a spectrum of formalisms such as variants of temporal logics (Pnueli 1992) suitable for this style. These logic-based formalisms provide a good mechanical support for checking the above two properties (though the general problem is undecidable).

#### 4.2 What's the concern of system specification?

As mentioned already, the major concern of this stage is to arrive at a model satisfying the requirements. That is, this stage provides an abstract model relative to an environment satisfying the requirements. As it provides a model, it is necessary for this stage to be concerned with abstract implementation operators such as non-determinism, concurrent and hierarchical that are concerned with the architecture the user has in mind for building the system. In other words, it is desirable that the specification is executable and simulatable from which further custom-made implementations can be derived. By definition consistency does not play a role except that one would have to ensure that there are non-vacuous domains of answers for the system. It must be noted that it is not necessary that there should be one-to-one correspondence between the requirements and the logical units of the system. The property of completeness also does not play a role except at successive refinements of the system specifications where some units may be omitted.

Traditionally, system specifications have been expressed as interacting state machines; such an approach inevitably suffers from over specification as the state machines represent an implementation. If the application is such that only one implementation is envisaged, an implementation oriented specification may be acceptable; but other applications, for example communication protocol specifications, envisage many distinct implementations. Formalisms such as CCS, CSP (Hoare & Jones 1978), (Berry 1992) are some of the formalisms that have been used as successful formalisms for system specifications.

#### 4.3 What does verifying specifications mean?

The notion of verification of specifications largely depends on the language frameworks used for writing specifications. It is usual to write specifications either in the axiomatic or the algebraic frameworks. Frameworks used for writing specifications can be broadly categorized into (Pnueli 1986):

1. Two language framework: This corresponds to the usual axiomatic specification wherein one language is used for specifying the properties of the program and another language is used for specifying the abstract model/the program itself.
2. Single language framework: Here, the same language is used for specifying the properties as well as the model or the implementation.

*Verification in two language framework:* One of the widely used methods of capturing the properties of the program are through state changes (Lampert 1983). As usual hierarchical specifications forms the basis of mastering complexity. Thus, the main question is to ensure that the *lower level specification* is a refinement of the *immediate higher level* in an iterative manner. To make it more precise, consider a specification at the lower level, say  $S_i$  described as

$S_i$  : there exist state functions  $g_1, \dots, g_s$  such that  $B_1, \dots, B_n$

and a higher level specification  $S_{i-1}$  given by,

$S_{i-1}$  : there exist state functions  $f_1, \dots, f_r$  such that  $A_1, \dots, A_m$

To show that the lower level specification is a refinement of the higher level, we must find expressions  $F_i(g_1, \dots, g_s)$  such that each  $A'_k$  obtained from  $A_k$  by performing substitutions

$$f_i \leftarrow F_i(g_1, \dots, g_s)$$

follows logically from the axioms  $B_1, \dots, B_n$ . Now consider a still lower level specification

$$S_{i+1} : \text{there exist state functions } h_1, \dots, h_t \text{ such that } C_1, \dots, C_p$$

This can be shown to be a correct refinement of  $S_i$  by finding substitutions

$$g_j \leftarrow G_j(h_1, \dots, h_t)$$

which yield formulae  $B'_i$  that can be proved from  $C_1, \dots, C_p$ . It follows that the formulae  $A''_k$  obtained from  $A_k$  by the substitutions

$$f_i \leftarrow F_i(G_1(h_1, \dots, h_t), \dots, G_s(h_1, \dots, h_t))$$

can be proved from  $B'_1, \dots, B'_n$ , which in turn can be proved from  $C_1, \dots, C_p$ . Thus, by transitivity, it follows that  $S_{i+1}$  is also a refinement of  $S_{i-1}$ . Hence, if each layer is correct, one can show that the lowest level of specification implements the highest level of specification.

The crux of saying that the *refinement is correct* becomes simple (with little leaps of intuition) when both the levels are assertions about the same model as *state refinements in that case can be structured appropriately to preserve the intended intuitions*. In case, the model on which assertions are made in the sequence of refinements are not close to machine languages, then one has to establish that *there is a correct implementation of the language*. It is here that a sound and a complete semantics of the language plays a vital role. It is possible to use variants of temporal logic (Lamport 1983) as the vehicle for specification (translation) at every stage of refinement and to go down to the levels of concrete high level programming languages or even lower level programming languages. For this purpose, one could use the mechanical theorem provers to advantage. Thus, establishment of the correctness of refinement corresponds to proving that the lowest level of specification implements every level in the hierarchy.

*Verification in single language framework:* The basis of verification in this framework is *equivalence of programs or specifications*. That is, one usually specifies a simpler program and shows that the two programs are behaviourally equivalent. The equivalence establishes that the implemented program behaves like the other program; that is, it is as good or as bad as the simpler program. Thus, a refinement relation  $\sqsupseteq$  can be captured for a hierarchical sequence  $P_0 \sqsupseteq P_1, \dots \sqsupseteq P_m$  moving from the simplest naive specification to a concrete implementation  $P_m$ .

#### 4.4 What style to choose?

It is evident from the above discussion that the two styles are complementary to each other and both stages play a vital role in the design cycle. In fact, the two stages play an important role irrespective of the language frameworks. In the context



of two-language frameworks the distinction is apparent. In the context of single-language frameworks the use can be seen if one visualizes the coarse level specifications/programs with which the fine-grain solution will be checked for equivalence. For the success of the method of two stages it is important to choose an appropriate specification language for each stage. Let us now take a brief look at the choice of specification languages.

The requirement specification essentially states *what system* is to be developed at *what costs*, and under *what constraints*. In other words, it is used

1. as a vehicle of communication,
2. as a scope for modifiability,
3. to constrain target systems, and
4. for accepting or rejecting final products.

Thus, it is necessary that any candidate specification language for this purpose should satisfy these requirements in general. From the point of view of (1)-(2), it is clear that the language must be understandable and modifiable; the latter property asks for the *conjunctive property*. From the point of view of (3), it follows that the language must be precise, unambiguous, internally consistent, and complete. Furthermore, the requirements specification should be minimal, i.e., define the smallest set of properties that will satisfy the users and originators. Otherwise, the specification may over-constrain the target system, so that some of the best solutions to design problems are unnecessarily excluded. For instance, at this level there is no need for the specification of nondeterminism and parallelism. Property (4) enforces the requirement of formal manipulatability (if verification is to be used) or testability (if testing is to be used) on the specification language.

Linear temporal-logic based formalisms (Pnueli 1986) are conjunctive and satisfy the needs for specifying the needed safety and eventuality properties to a large extent. However, as temporal logic suffers from its orientation towards eventuality rather than immediacy or quantization, one would have to use real-time temporal logic formalisms. Another important consideration in the selection of the particular variant has been the availability of model checkers for the language on hand so that one can attempt at automatic verification of properties of finite state systems.

For system specification purposes, abstract notations such as CCS, temporal logic of actions, refinements of temporal logic and concrete languages such as CSP, Esterel etc. are applicable. The choice will also depend upon how close the specification is to be with respect to the implementation and also the type of verification and the tools one would be interested for verification purposes.

#### 4.5 Software maintenance and formal methods

The need for modifications to old programs cannot be under estimated. Some of the changes are necessitated by the need to adapt the program onto another system, the need to arrive at a new system with some changes (perhaps incremental) in the specification of some of the components or discovery of errors. Often it is not possible to write an entirely new program due to considerations of time and cost. It is here the choice of the formalism, documentation and the ingenuity of the programmer plays a crucial role. If the formalism chosen is *compositional (or modular)* then the

task would be proportional to the changes required. At least if the programmer had properly articulated (and documented!) those aspects of the program that could perhaps be subjected to future modifications, then the task becomes manageable. In fact, modifications indeed could be effected economically and reliably if one had chosen proper mathematical (formal) method of development (Hoare & Jones 1986). If the structure of the program  $P$  is the same as the structure of the specification  $S$ , then it is sufficient to ensure that the modified component meets the modified specification. But it is not always possible to preserve the structure of specification in the design of the program. This is so because a specification is often most clearly structured with the aid of such logical operators such as negation and conjunction, which are not available in an implemented programming language. Nevertheless, mathematics can come to rescue. If the program  $P$  has an *approximate inverse*  $P^{-1}$ , defined in the same way as for the *quotient*, then it is possible to calculate the new proof obligation of the program. For details the reader is referred to (Hoare 1989).

#### 4.6 Validation

Correctness is the process of establishing either the equivalence of two objects or refinability of one object from the other. Thus, we can say that a system or an implementation is correct if it meets the *designers' intentions*. Thus, unless the the designers' intentions are formalized, the notion of correctness becomes vacuous. The notions of verification in the single- and two-language frameworks have been already discussed in the previous sections. One of the important processes in the development of a system is to show that the system indeed satisfies the *Designer's intentions*. This process is often referred to as *validation*. It must be noted that *Designers' intentions* is not completely formal. Thus, validation is not a complete *formal process*. Thus, for validating a system, one uses test scenarios and uses simulation etc. For the latter, if the specification is executable (like system specifications) then one can validate the system easily. It is important to note that the process of validation is an informal process; the purpose is to ensure that the specification conforms to the intents.

### 5. Case studies of formal methods in the development of safety critical systems

In the following, we take a brief look at some representative industrial scale software systems for which formal methods have been applied for design, validation, review of the system. We briefly discuss the experience in the use of formal methods in the design, development or review of large software systems in the context of the following areas:

1. Railway Signaling Systems.
2. Nuclear Power Plants.
3. Aviation Control.
4. Medical Systems.

The first example illustrates the study of the development of Paris Metro Signaling system where formal methods were deployed right from the beginning and resulted in substantial economic benefits. The second example illustrates the deployment of formal methods for gaining the level of confidence in an already developed system. The third example illustrates how a simple formal model aids in the understanding of the design by engineers and users as well and thus, enables to remove inconsistencies. The last example illustrates the need for a formal analysis of systems particularly medical life-critical systems for gaining the confidence of regulatory agencies and users alike. Many other case studies have been reported by Gerhart *et al* (1994).

### 5.1 Paris Metro signaling systems

In the Paris Metro signaling system, formal analysis was used to determine that the separation time between two trains on the Paris Metro system could be reduced from 2 minutes 30 seconds to 2 minutes while maintaining safety requirements. More importantly, the successful deployment of the signaling system removed the need for building a new third railway line, thus saving billions of dollars in costs. There was a serious concern about the predictability and reliability of large amount of software and the associated cost in testing the system exhaustively because of safety considerations. This case study is a positive example of the major benefits to be gained by the deployment of formal methods throughout the design and development of the system.

The developers started with Hoare's method of proving correctness of programs (note that the development of the system started in early 80's!). Using this technique three sets of proven software were developed. The question that arose was: *Are these sets consistent?* Consistency of validated sets of software were approached as follows:

1. Top-to-bottom re-specification and refinement.
2. Bottom-to-top re-specification and verification.
3. Match the two at some level close to the code.

Such a revalidation by proof re-engineering validated both their method and design. The tools used in the design and validation (revalidation) of the system were broadly the verification-condition generators and the tools around the B Method (Abrial *et al* 1991).

The success of the project was not due to a particular methodology for formal methods used or the advanced nature of tools deployed. Such methods and approaches are supported in most verification systems, and more advanced and powerful theorem provers with sophisticated interfaces are already available. The important message learnt was the value of combining formal methods with prototyping, simulation and testing. From the point of view of the use of formal methods in the development of software, the experience shows:

1. Formal specifications help developers understand the requirements clearly.
2. Formal methods play a major role in developing confidence in the system.

The experience of the case study has been reused by the GEC Alsthom Transport Corporation to sell similar related technology and expertise in other projects such

as SNCF, France and Controlled Deceleration Control, Calcutta, India. In these projects, the application of formal methods is viewed as a success, and the resulting software is considered to be efficient and of good quality. Formal methods have now become a part of the GEC Alsthom's software strategy.

## 5.2 *Darlington Nuclear Generating Station*

In contrast to the Paris Metro signaling system in which formal methods were used in all stages of the design and development, the Darlington case study is an example to show how the formal methods helped in enhancing confidence in the functionality of a system that was developed without using any formal methods. This suggests that formal methods can be useful in analyzing an existing design and implementation, a kind of reverse engineering.

The Darlington station is a four-reactor nuclear plant in Toronto, Canada. The reactor had two fully computerized shutdown systems; the first one drops neutron-absorbing rods into the core; the second injects liquid poison into the moderator. The systems are safety-critical and require high levels of confidence. The shutdown system had been designed and developed using conventional software engineering practices. It had gone through unit testing, integration testing, validation testing, in-site trajectory-based random testing, software assessment, and software hazard analysis. However, the review of the software before licensing uncovered discrepancies and raised doubts as to whether the software implemented the requirements correctly. There were also some ambiguities and doubts about the requirements themselves. Since this was the first such system using computerized shutdown system, the Atomic Energy Control Board of Canada was not willing to grant the license to operate the plant because of these doubts about the reliability of the shutdown systems.

Each month's delay in getting a license was costing Ontario Hydro \$20 million in interest payments for the whole nuclear generating station. The Ontario Hydro was not keen on redesigning the software, and in order to reduce the licensing impasse, a compromise was reached. The shutdown software was to be formally inspected:

1. Formalize informal requirements by generating specification tables.
2. Use the existing code to develop program-function tables for it.
3. Demonstrate that the code is consistent with the specifications.

In other words, the approach was primarily a reverse-engineering exercise, in the sense that code already existed when the formal specifications were written. The method called *Software Cost Reduction* (SCR) originally developed at the US Naval Research Labs was adopted for the inspection.

The above three steps were done by three independent teams. Deriving specifications from informal requirements consisted of arriving at mathematical formulas which would show the effect anticipated; deriving program-function tables corresponded to arriving the effect of code procedures. Most of this analysis was done by hand; proofs were done by hand with little use of automated tools. Microsoft Excel was used to produce these tables. Finally, the proof of consistency consisted of manually comparing the specification and program-function tables, and transforming the tables. The last team also reported the discrepancies to the other two teams.

The experience of gaining confidence through such an approach was felt worthy by the developers. The main lessons learnt can be summarized as follows:

1. The analysis of requirements clarified the specifications and removed some safety-related ambiguities. In particular, it became easier to understand the operational behavior.
2. Without the formal methods, there would have been no proofs of correspondence between specification and program-function tables. The formalisms became handy in the review extensively and raised the level of the confidence of the system. It also helped identify some problems with timings.

The result of this exercise was that the Atomic Energy Control Board of Canada felt more confident in granting operational license for the plant.

### 5.3 Aviation control systems

The third case study concerns the Traffic Alert and Collision Avoidance (TCAS) system being developed by the US Federal Aviation Administration (FAA). All aircrafts having more than 30 seats are mandated to have such a system installed to avoid midair collisions. The goal has been to develop requirements and design of TCAS Logic and TCAS Surveillance, which is to provide traffic advisories and recommend vertical maneuvers in the event of an impending collision.

There existed a natural language specification of TCAS which included a pseudo-code description of CAS Logic. In 1990, FAA and others had become concerned about the informal requirements specifications especially because errors had been uncovered through the simulation of the pseudo-code; the specification had to be revised quite a few times. FAA wanted to further clarify the TCAS requirements and obtained improved confidence in the system.

The TCAS methodology is directed at process-control systems designed to maintain an acceptable relationship between a system's inputs and outputs in the presence of disturbances to the process. Leveson (1986) and her students at University of California, Irvine, have developed a notation based on Harel's Statecharts for specifying such systems. At the beginning, they wrote most of the specifications of TCAS, and domain experts reviewed them, but later the domain experts took over. The experience shows the the Statechart-like specification has made it easy to specify the transition logic of the TCAS subsystems and has provided a good satisfaction and a comprehension of the specifications for the designers. Engineers and others not trained in formal methods can review and modify the specification. In particular, the use of formal methods eliminated the development of an informal natural language specification.

### 5.4 Medical therapy systems

The increasing use of medical instruments for life-support and life-critical systems has made it necessary to ensure a high degree of dependability/reliability. In fact the regulatory requirements from agencies like US Food and Drug Administration has made it mandatory for software developers to establish the safety not only of the software being designed but also establish the safety of the software that has already been developed.

In (Mojdehbakhsh *et al* 1994), a retrofitting software safety analysis is developed for implantable cardiac-rhythm-management systems. The developers and the customers were convinced that the formal analysis successfully identified and mitigated numerous software-safety faults and eliminated several hazards. Some of the safety faults that had been undetected by the reliability analysis (essentially, done through the system fault-tree approach) has been captured through such an analysis.

The use of formal specifications (Ladeau & Freeman 1991) in the development of a bedside instrument used to monitor vital signs in patients in intensive care units and operating rooms has shown the uncovering of several problems and ambiguities in the informal product specification. Several other successful applications have been summarized by Bowen & Stavridou (1993).

## 6. What have we learned from the use of formal methods?

In this section, we briefly discuss the feedback from the experiences in using formal methods in the development of safety critical systems (Kapur & Shyamasundar 1994). The use of formal methods is pivotal in the design, development, and maintenance of complex systems. They also form a good medium for review and reverse engineering. However formal methods cannot and should not be viewed as a substitute for identification and use of good abstractions relevant for the application domain. It is the proper use of appropriate abstractions that leads to good design, structure and implementation of complex systems. Selection and deployment of relevant abstractions can also reduce the cost and increase the effectiveness of formal methods. In fact, formal methods can be used to assess the selection of appropriate abstractions; a simple case of specifying transitions in the TCAS using predicate calculus has made the specifications more comprehensible to engineers and non-engineers.

Formal methods are likely to be most cost-effective and have a bigger payoff when used in the earlier stages of the system life-cycle. In the later stages of the life-cycle, the use of formal methods is expensive and time-consuming. It is thus more economical to apply formal methods to specifications and designs than to programs. There are at least two reasons. Firstly, sooner the mistakes and errors are found in a specification, better it is in terms of cost effectiveness and avoiding delays in completing a software project. According to Fairly (1985), it is 5 times more costly to correct a requirement fault at the design stage than during the initial requirements, 20 to 50 times more costly to correct it at acceptance testing, and 100 to 200 times more costly to correct the problem once the system is in operation.

Studies show that almost all accidents involving computerized process-control systems are due to inadequate design foresight and requirement specifications, including incomplete or wrong assumptions about the behaviour or operation of the controlled systems, and unanticipated states of the controlled system and its environment (Leveson 1991). In 203 formal inspections of six projects at JPL, Kelly, Sherif and Hops found that requirement documents averaged one major defect every three pages (a page is 38 lines of text) compared with one every 20 pages for code. Two-thirds of defects in requirements were omissions (Rushby 1993). Similar experience is reported from the space-shuttle, Voyager and Galileo-projects at NASA. The need for reliable, correct specification cannot be overemphasized.

Secondly, specifications and designs are high-level and can be described function-

ally where formal methods can be applied nicely and efficiently. There is a need for testing specifications to enhance designer's confidence in them and ensuring that specifications have the desired structural properties. Concentrating on consistency and completeness at higher levels of specification goes a long way in the realization of quality software. It is to be noted that several automated tools of deduction are indeed available for this purpose presently.

It was increasingly felt in most case studies that formal requirement and design specifications are essential. They not only assist in early discovery of incomplete and ambiguous specifications, they lead to clarification, crystallize incompatibilities in understanding of the clients and designers thus resulting in better comprehension of client's informal requirements and gains in assurance. The recent work in the context of TCAS and Paris Metro Signaling System is an indication of that.

A new programming paradigm called *synchronous paradigm* (Berry & Gonthier 1992) has provided a good basis for programming reactive systems; some of the languages based on this paradigm are Esterel, Lustre, Signal, and Statecharts. Esterel, one of the earliest languages of this family, has a highly sophisticated environment for design and verification. A striking feature of this system is that the the verified program and executed program are almost the same. Such a property makes it possible to have high levels of confidence in the verification done through the tools based on *bisimulation*. This paradigm will be further explored in the next section.

Formal methods are somewhat more difficult and cumbersome to apply to programs because of having to deal with state and state change. At this stage, one can rely on testing, validation tools and function/procedure tables. Program verification is most advantageous for (i) safety-critical kernels whose ultra-reliability is crucial for the behavior of the rest of the system, and (ii) for hardware chips whose behavior can be described functionally. For an overview of the use of theorem proving in software design and related issues, see (Kapur *et al* (1992) and Kapur (1993).

It should be possible to integrate formal methods to existing software engineering practices in a smooth manner thus enhancing the confidence in software engineering methodologies. Formal methods are most productive and effective when they are combined with other techniques such as prototyping, simulation, verification and testing.

Although having access to automated tools for using formal methods such as theorem provers, simplifiers, etc., would be handy, most studies seem to suggest that formal methods can be used without much automated support. In fact, in most industrial case studies, automated support was available mostly for type checking and syntactic correctness. Given that theorem proving technology has substantially developed over the past two decades, integration of powerful theorem provers and proof checkers could go a long way in mitigating some of the work involved in doing proofs. In the Darlington case study discussed earlier, all tasks (except for documentation) were performed manually, because of which the process of formally inspecting the already developed software became very labor intensive; it was estimated that 35 person-years of effort was expended on the walk-throughs! If there were tools available to perform automated reasoning for walk-throughs as well as to record dependencies among programs through their function tables, the time taken to perform walk-throughs would have been considerably reduced thus saving considerable money. For a discussion of how theorem proving technology should be further advanced for making it better suitable for software design, see Kapur (1993).

Regulatory agencies also appear to be satisfied with the use of formal methods for gaining assurance in product development even if it is not developed using formal methods. This is a kind of reverse engineering where formal methods are used to assure the regulators that software did what it was supposed to do.

1993, 1994) and

Re-usability is critical for cost-effectiveness and acceptance of formal methods. In the case of the Paris Metro project, the reuse of models and theories was helpful in two subsequent projects at GEC Alsthom. GEC Alsthom not only used formal methods to establish that there was no need for an extra railway line, but the developed expertise is being used to market its services in other projects. Object-oriented methodology can be helpful in designing reusable code, designs, theories, proofs, as well as specifications.

## 7. Synchronous paradigm for reactive systems

In this section, we highlight one of the new paradigms referred to as the *synchronous paradigm* which is being used successfully in the design of reactive systems and hardware systems. The synchronous paradigm is founded on the *perfect synchrony hypothesis*. The hypothesis states that the considered reactive programs respond in no time (i.e., the elapsed time is not observable) and produce their outputs *synchronously* with their inputs. In fact, this hypothesis takes roots in classical mechanics, being akin to the Newtonian's instantaneous body interaction principle – which is still useful in practice even though it has limitations at speeds beyond some ranges<sup>4</sup>. Further, perfect synchrony hypothesis can be seen through the assumption of an ideal perfect real-time machine in the works on control theory. Esterel is the first programming language to have been based on this principle.

First, let us take a quick look at the immediate gains for the specification of real-time systems with such an approach. One of the immediate gains of the perfect synchrony hypothesis can be seen by looking at the following paradoxical question in asynchronous languages, such as Ada, supporting specification of real-time:

Is delay 8 ; delay 6  $\equiv$  delay 14 ?

There is no clear cut answer to this question in an asynchronous framework. In the semantic theories of CSP-R (Koymans *et al* 1988), and Timed-CSP (PRG 1992), the paradox has been overcome through the notion of executional models like *maximal parallelism* and *maximal progress* respectively. Obviously, such a paradox clearly does not exist under the perfect synchrony hypothesis.

The next most important gain is that the notion of *physical time* can be replaced by a simple notion of *order* among events; the only relevant notions are the simultaneity and precedences among events. Thus, physical time does not play any special role; it will be handled as an external event, exactly as any other event coming from the programming environment. This is referred to as the *multiform time*. As an example, let us consider the two following requirements:

<sup>4</sup>It may be recalled that Einsteinian physics does not void the utility of Newtonian physics. Both are useful approximations that may be beneficially used for modeling and analysis of systems that fall into well defined ranges, where the main distinguishing parameter is how close are the typical speeds in the system to the speed of light. The same observation holds for the discipline of application of formal methods for the specification, design and verification of reliable reactive systems.



The train must stop within 10 seconds

The train must stop within 100 metres

Conceptually, these two requirements are the same; one could argue as to which is specific or which is more general. In systems having an internal clock and handled by special statements, conceptually similar requirements will be expressed in different ways. However, in languages following the above hypothesis, there is no notion of internal clock and hence, will be expressed by similar precedence constraints. In other words, a reactive system is completely event driven and we can consider the life (history) of a system to be divided into *instants* that are the moments where it reacts; or the history of a system is a totally ordered sequence of logical instants. Thus, one can speak of the *i*th instant of a program. Event occurrences which happen at the same logical instant are considered simultaneous, those which happen at different instants are ordered as their instants of occurrences. Apart from these logical instants, nothing happens either in the system or in its environment. Finally, all the processes of the system have the same knowledge of the events occurring at a given instant.

First, let us see informally how we can specify a vending machine. The behaviour of a typical Biscuit Machine can be described as follows:

1. It waits for insertion of 5 cents, ejecting a packet of biscuits or attempting to eject when there are no packets left,
2. The serviceman may open the machine by inserting the service key. After being re-stocked and the removal of the service key, the vending machine resumes where execution was interrupted.
3. The above process is repeated except if the power goes down, in which case it is possible to be cheated of 5 cents.

First, let us describe the basic behaviour considering only requirement (1) assuming that there is enough stock of biscuits and there is no power tripping. In this case, the observable events are waiting for 5 cents and ejecting the pack of biscuits after receiving 5 cents.

```

loop
  await ?5_CENTS; (* wait for 5_cents *)
  if AVAILABLE then "EJECT"
forever

```

The synchrony hypothesis avoids questions such as: What happens if one puts several 5 cents since for an input there is a *deterministic instantaneous* reaction.

Considering requirement (2), we can describe the behaviour by

```

loop
  await ?5_CENTS;
  if AVAILABLE then "EJECT";
  else
    await INSERT_SERVICE_KEY;
    "EJECT";
  endif
forever

```

It is easy to observe that if the local actions are assumed to be instantaneous then the above description gives the intended behaviour; note that *instantaneous* does

not mean that order of events in the same instant can be ignored; in fact, the order of events should necessarily be preserved- the only thing is that time separation between events is not observable. For taking into account the third requirement, we need some sort of a preemption. This can be done by introducing a guard to watch the execution. The behaviour can now be described by,

```

loop
  do
    (* Begin of a guarded statement *)
    await ?5.CENTS;
    if AVAILABLE then "EJECT";
    else
      await INSERT_SERVICE_KEY;
      "EJECT";
    endif
    watching ?powerdown      (* GUARD *)
  end
forever

```

In the above code, *do stat watching S* corresponds to preemption of the execution of the it body on receiving S. It can be observed that in the above example, we have been able to develop the program by taking requirements one at a time (that is, conjunctive). In fact, it appears that this feature holds in general. Informally, one can see that there is an in built priority in the specification or interrupts in the implementation. Later, one can see that Esterel framework provides for a nice integration of interrupts and exceptions.

For a language to support reactive specification, we can see that it should support broadly the following features:

- **Sound mathematical semantics:** The very need of safety requirements of the applications are enough to convince oneself that the behaviour of a program must have a sound and a unique meaning. Such a requirement calls for a sound mathematical semantics.
- **Determinism:** In any reactive system, one should be able to predict the behaviour relative to the sequence of sets of events. Obviously, this is a must. One should carefully distinguish between the need of nondeterminism in specification and the requirement of deterministic behaviour with respect to possible set/sequence of input stimuli.
- **Concurrency:** Most applications involve naturally concurrent and communicating components and thus, the language should support its specification at a logical level. It must however be pointed out that one should carefully distinguish between compile-time concurrency and physical run-time concurrency. It is of interest to note that logical concurrency makes it easier to write programs but does not necessarily correspond to the architecture of the executing machine. For an interesting discussion on parallelism as a structuring technique, the reader is referred to [MuSe 92]. However, the physical concurrency reflects the underlying machine's architecture. In the context of distributed programs, physical parallelism is often associated with code *distribution* that correspond to execute some sub-programs on distinct machines communicating through a network.

- Verification: From the reliability point of view, it is absolutely essential that it should be possible to verify the properties of the programs. Since the task of verification is quite complex, it is necessary to have automatic tools to assist the verification of the program.

### 7.1 Languages, validation and verification

There are various languages that support synchronous paradigm. The principal languages are Esterel (imperative), Lustre (data flow), Signal (equational), Statecharts (graphical) etc. Esterel and Lustre are synchronous, and deterministic where as Statecharts is not necessarily deterministic; Signal is not completely reactive (due to the oversampling operator). Esterel is one of the earliest languages of this family and supports a powerful programming environment. The main tools can be classified into:

1. Simulation and development tools: These tools execute compiled automata of Esterel programs instants by instants. It can be linked with a *C* standard library that allows interactive simulation through the keyboard. It also supports a graphic simulation through the X-window library. Recently, it supports an integrated environment called AGEL. The advantage of these tools is that one can validate the system and visualize the working of the program (and the reactive modules) clearly.
2. Verification tools: The Esterel program is compiled into an automata. Now, one can use various techniques for verifying the finite-state automata and also the tools that are used for verification using process calculi. The tools such as AUTO/AUTOGRAPH used for verification of process calculi have been integrated in the environment and thus, enables one to verify using notions such as bisimulation and observable criteria.

One of the main advantages of the system is that the verified code is close to implementation and hence, the proof becomes very reliable. In fact, it is because of this reason, the designer of Esterel, G. Berry claims *What You Prove Is What You Execute*. The language is being widely used for the development of reactive systems and the synthesis of hardware.

### 7.2 Illustrative specifications

In this section, we illustrate the development of an Esterel program from a typical specification; we only give fragments of code for lack of space and use loose syntax.

#### Slotted ALOHA

This example illustrates how the features such as *broadcast* and *logical concurrency* are helpful in arriving at concrete implementations.

In this protocol, the satellite acts as a repeater, rebroadcasting messages received from independent ground stations. The principal features of the protocol (Tanenbaum 1981) are captured below:

1. The satellite broadcasts a clock signal dividing time into discrete intervals, or slots; it is this feature that tries to avoid collisions.

2. A ground station which is ready to transmit must wait for the next slot before broadcasting.
3. If only one station transmits during a particular slot, then the satellite will receive the message in tact, and will rebroadcast it to all ground stations. If two users broadcast simultaneously, the satellite will rebroadcast the sum of two incoming signals, resulting in garbage. It is assumed that each packet contains a checksum strong enough to permit the receiver to detect all collisions, so damaged packets can be discarded.
4. The transmission is in terms of packets of the same length (the throughput increases with such restrictions).
5. Another important property of satellite packet broadcasting is that the sender can listen for his own packet, *one round-trip time after sending it*. Since the sender can tell from the bounced message whether or not a collision has occurred, there is no need for explicit destination to source acknowledgements. If the packet was garbled, the sender learns of the problem simultaneously with the receiver and can take appropriate action without having to be told.

In the following, we describe an Esterel development of the specification; for purposes of brevity we describe the main aspect of the specification leaving the formal declarations of events and signals.

The satellite can be modelled as the parallel combination of two processes `CLOCK` and `REFLECT`,

```
SATELLITE :: CLOCK || REFLECT
```

The `CLOCK` process emits a PIP signal every TM time units (or ticks).

```
CLOCK :: loop
    await TM ticks;
    emit PIP;
  forever
```

The `REFLECT` process is able to receive packets on `CHAN_UP` and broadcast the messages on `CHAN_DOWN`. Collisions can happen on the packets being received on `CHAN_UP`. It is ready to receive packets in every slot (after a PIP is emitted by the `CLOCK` process). In the specification given below, we have used `APP` to denote the append operation which appends an element to a list and `HD` denotes the operator `HEAD` of a list. Let `m` be the size of the packet.

```
REFLECT :: loop
    await PIP;
    PKT:=EMPTY;
    repeat m times
    do
        await CHAN_UP;
        APP(PKT, ?CHAN_UP) ;
    and repeat:
```

```

    end repeat;
  forever

```

The GROUND\_STATION is able to send a packet of fixed size during every slot on CHAN\_UP. After sending the packet it awaits for the reception of the same on CHAN\_DOWN. If the message received is garbled due to collisions ( which can be detected by the *strong checksum* property), then the GROUND\_STATION retransmits; this is repeated till a successful transmission takes place. Note that the input signal START also aids in avoiding collisions.

```

GROUND_STATION :: RETRANSMIT:=FALSE;
  loop
    await PIP; (* await for the clock signal from the clock *)
    await START;
    (* await for the input signal at the ground_station
       to instantiate transmission of a packet *)
    if not (RETRANSMIT) then
      CNT_ PKT:= NEW_ PKT
    else CNT_ PKT:= OLD_ PKT
    RETRANSMIT:=FALSE;
    i:=1;
    repeat m times (* m is the size of the packet *)
      do
        [emit CHAN_UP(CNT_ PKT[I]) || await tick];
        I:=I+1;
      end;
      trap T in
        do
          [await CHAN_DOWN || await tick];
          if NOT_VALID(?CHAN_DOWN) then
            begin RETRANSMIT:=TRUE;
              exit T;
            end;
          watching m ticks (* Receiving continues till the complete packet
                             is received or it knows that the message is garbled *)
        end
      end trap
    forever

```

### 7.3 Application hybrid and timed systems

Programming languages based on the *perfect synchrony paradigm* have proven useful for programming reactive systems. One of the main reasons for its success is that it permits the programmer to focus on the logic of reactions and makes it possible to use several automata-based verification systems for correctness proofs. Further, the correctness proofs of programs follow their implementation very closely and hence, are more robust and reliable. However, the application is limited to clocked systems. If we look at complex reactive system specifications such as process control systems and robotic applications, the need for features such as (i) asynchronous events (events that can happen arbitrarily close to each other), (ii) integration of discrete and continuous components (continuous components may cause continuous change in the values of some state variables according to some physical law), and (iii) explicit clock times, become apparent.

Further, the use of *hybrid systems* (Kesten & Pnueli 1992) is becoming very extensive. Hybrid systems are systems that combine discrete and continuous computations. Hybrid system model contains activities that modify their variables continuously over intervals of positive duration in addition to the familiar transitions that change the values of variables instantaneously, representing the discrete components. Many systems that interact with a physical environment such as a digital module controlling a process or a manufacturing plant, a digital-analog guidance of transport systems, control of a robot, flexible manufacturing systems etc., can benefit from the study of hybrid models.

In (Berry *et al* 1993), a paradigm referred to as *Communicating reactive processes* (CRP) that provides a unification of perfect synchrony and asynchrony has been presented. In (Shyamasundar 1993), an extension of CRP, referred to as *Timed CRP* has been envisaged that

- models *continuous* computations and thus, provides a convenient formalism for specifying hybrid systems, and
- models asynchronous systems operating in dense real-time domains.

Let us now consider the basic idea for specifying hybrid systems that combine discrete and continuous components possibly with the need to reference clocks explicitly. For consistency, it is necessary to have a consistent assumption about the progress of the computation as the system evolves. In the timed CRP, it has been possible to provide such a consistent assumption through the clocked semantics of CRP and the interpretation of the clocks in terms of the *exec* primitive. One of the interesting features is that hybrid systems without explicit references to clocks can be specified by a subset of timed CRP consisting of just the statements of *ESTEREL* and the *exec* primitive. The full version of timed CRP can specify dense asynchronous systems with explicit references to clocks (cf. (Shyamasundar 1994)).

## 8. Role of safety in the design of safety-critical systems

With the advances in technology, computers are being increasingly used to monitor and/or control complex time-critical physical processes or mechanical devices where a run-time error or failure could result in loss of property, injury or even death. Such systems are usually termed *safety-critical* systems.

Safety is associated with the notion of risks such as loss of property, injury, death or damage to the environment. In other words, safety requirements are concerned with making the system *mishap-free* whereas reliability is concerned with making the system failure-free (Leveson 1986, 1991). Software safety is part of system-safety (Leveson 1991). Ensuring system-safety involves:

1. Identifying hazards and assessing the risks involved.
2. Designing ways to avoid or control them.

Thus, it is essential to arrive at system-fault tree from which one has to arrive at safe-system keeping in mind that all the hazards cannot be completely eliminated.

### 8.1 *How much is safety worth?*

An increasing number of computerized safety-critical systems are currently being deployed in such areas as transportation and nuclear power production, or will be largely deployed tomorrow in medical computing, automotive electronics, etc. Critical applications such as nuclear control systems, flight control systems, life-support systems have extreme safety requirements. For instance, FAA and NASA have established a requirement of less than  $10^{-10}$  safety-critical failures per hour throughout a 10-hour flight, a level roughly equivalent to one failure per million years of operation.

For hardware component faults, it is possible to achieve these low failure rates by use of highly reliable microelectronics, together with replication and adaptive majority voting. The primary factors contributing to unreliable operation are design faults, possibly in the hardware but more probably in the software. Software faults present the greater risk of system failure, because only relatively simple functions are mechanized in hardware while the most complex parts of a systems are implemented in hardware. The statistical evidence that software is the current bottleneck in achieving dependability of Information and Communication Technologies, together with the recognition that probabilistic assessment of software reliability to levels commensurate with safety requirements (e.g.  $10^9$ /h or  $10^5$  per demand) is currently out of reach, has led to highly labor intensive approaches for the development and validation of operational safety-critical software. Be they undertaken via traditional software engineering approaches or via mathematically formal approaches, orders of magnitudes of effort dedicated to the development and validation of such software are in the range of 10 man-years per 1000 lines of code, for software ranging from a few thousands to a few tens of thousands lines of code.

As the pervasiveness of software induces a clear tendency to complexifying the functions it is expected to fulfill, producing dependable software for critical applications of sustainable cost requires the identification and formulation of abstractions which are at the same time rigorous and representative of both the informatics constructs and the environment where the corresponding computing systems are intended to operate. At the same time, recent unfortunate examples have shown that computerized systems which were not initially felt to be safety-critical, and thus not built according to high costly standards, have endangered human lives upon failures, be they a relatively modestly complex apparatus such as the Therac 25 radiation therapy system, or distributed systems such as the communication system of the London Ambulance Service. Moreover, nation-wide failures of large computing and communication systems which cannot be built at the above cost, such as the outage of the inter-city phone system or the Internet collapse in the USA, can have indirect safety-related consequences, be they caused by accidental or malicious events.

In the case of embedded systems, most of the times it is quite difficult to provide a behavioural decomposition even if one ignores the need for the decompositions to be the basis for system design. In other words the separation of concerns turns out to be extremely difficult. For example, even small real-time systems such as a tactical embedded system for an aircraft might be simultaneously maintaining a radar display, calculating weapon trajectories, performing navigation functions etc. In these kinds of systems, one sees that

- the code implementing the various tasks is mixed together such that it is

difficult to determine which task(s) a given part of the code performs, and

the timing dependencies between code sections are such that changing the timing characteristics of one section may affect whether or not many otherwise unrelated tasks meet their deadlines.

The question is how do we go about? As discussed already, a general strategy is to base the design on a formal method. Of course, at this stage, it is necessary to do the hazard analysis by various techniques such as (1) design reviews and walkthroughs, (2) fault tree analysis, (3) failure modes and effect analysis etc. Synchronous paradigm has shown a good promise and has various components such as dependable compilers, simulators, verifiers etc, it is natural that one should base the development around this paradigm. In fact, the tools available on such systems also aid the hazard analysis techniques mentioned above. One of the important factors to be kept in mind is that in the design of reactive systems one should first go about specializing for the class of systems one is concerned with rather than going about for the generalized system design. In the next section, we discuss some of the measures that should be taken in India that would aid in the development of reliable reactive systems and reliable software in general.

## 9. An approach towards development of quality software

In recent years, India has been leaping into hi-tech areas such as communication, transport, nuclear industry, military tactical systems. Directly and indirectly, there is a large involvement in the development of software for embedded systems. Some of the failures in such projects can be attributed to software faults. Some of the hi-tech projects/systems being planned in India are facing a series of problems due to the fear in the community that a failure in the system will endanger the society or will lead to environmental catastrophes. We would make a major headway in overcoming such bottlenecks if we can

- develop safety standards and regulations in the use of embedded safety-critical software,
- place mandate for the use of formal methods and languages with sound rigorous mathematical basis, and
- agree to apply reverse-engineering techniques for safety-critical systems and evaluate them; this would gain the confidence of the society and also the design errors can be corrected and further, catastrophes can be avoided.

In fact, several countries such as UK have promulgated such standards as a must. Arriving at such standards and regulations will go a long way in the development of science, and technology in India.

One of the foremost things of concern in the Indian context is the need to convince the industry of the use of formal methods. The use of formal methods need to be made cost-effective, which is possible. Related to this point is providing training to professionals in the use of the formal methods. It is here that a major effort is required by academics and software professionals in India.

It can be argued that a thrust towards the development of quality software will also have an enormous impact on the economy in the context of software export.



1. Given that manpower in India is much cheaper, it is our contention that strong economic arguments can be made for the use of formal methods for generating safety/quality software. Generation of safety/quality software using formal methods is smart-labor-intensive and not equipment or commercial software intensive. We believe that programmers and software engineers in India (especially from top technical institutions) typically have a stronger background in mathematics and analytic methods than their counterparts in the west, especially the US. They can be trained quickly and with lesser economic cost in the use of formal method and the associated tools.
2. For sophisticated machinery and equipment to be sold to other third world countries, India would be competing with Western countries. It may become essential to develop methods for reliable software. India may develop a slight edge over Western countries on this front because of cheaper cost in producing software due to cheaper labor costs. Also it is estimated that there would be a large demand based on the synchronous paradigm for the development of reactive systems and hardware systems. Hence, investment on the synchronous programming technology will not only aid in the development of systems but also will put India on an advantageous position in the software development (for consumption or export).
3. India can take a lead in developing tools and methodologies for generating ultra-reliable software. It is technically feasible to generate a next-generation environment for designing systems based on formal methods.

What should be key characteristics and features of a next-generation integrated environment for providing support for formal methods in the life-cycle of system design and development? Many technical issues related to integrating automated reasoning tools, simplifiers, specification analyzers are discussed by Kapur (1993). We would like to emphasize that formal methods should be integrated into the existing methods in a

- (a) Localized way: that is, it should not be the case that formal methods have to be applied to the whole system. It should be possible to test them and experiment with them on parts of a system.
- (b) Reversible way: If for some reason, the use of formal methods has to be abandoned that should be possible in graceful without causing significant delays in the development of the system.

In other words, the use of formal methods should not interfere with the existing development process, that is, the additional features required for deploying formal methods, if ignored, should not cause major disruption in the life-cycle. It is the careful integration of formal methods with existing methods that include prototyping, testing, structured walk-throughs, validation, hazard analysis, fault-tree analysis, and simulation, etc., which has led to reasonable success in the development of critical software and hardware in practice (Weber-Wulff 1993).

## References

- Abrial J R, Lee M K O, Nielson D S, Scharbach P N, Sorenson I H 1991 The B-method. *VDM 91. Lecture Notes in Computer Science* 552: 398-405
- Berry G 1992 A hardware implementation of pure Esterel. *Sadhana* 17: 95-139
- Berry G, Gonthier G 1988 The Esterel synchronous programming language: Design, semantics, implementation. *Rapport de Recherche 842, INRIA 1988, Science of Computer Programming* 19: 87-152
- Berry G, Ramesh S, Shyamasundar R K 1993 Communicating Reactive Processes. *20th ACM Symposium on Principles of Programming Languages* South Carolina, pp 85-99
- Bowen J, Stavridou V 1993 Safety-Critical Systems: formal methods and standards. *IEE/BCS Software Eng. J.* 8: 189-209
- Brinksma E 1992 What is the method in formal methods. *Formal Description Techniques* (eds) K R Parker, G A Rose (New York: Elsevier Science)
- Fairly R E 1985 *Software engineering concepts* (New York: McGraw-Hill)
- Gerhart S, Craigen D, Ralston T 1994 Experience with formal methods in critical systems. *IEEE Software* : 21-39
- Hoare C A R, Jones C B 1989 *Essays in computing science* (London: Prentice Hall International)
- Kapur D 1993 Automated reasoning in software design. *CSA Jubilee Workshop on Computing and Intelligent Systems* (New Delhi: Tata McGraw-Hill) pp 201-216
- Kapur D, Musser D R, Nie X 1992 The Tecton proof system. *Proc. of the Workshop on Formal Methods in Databases and Software Engineering, Workshop in Computing Series* (eds) Alagar, Lakshmanan, Sadri (Berlin: Springer-Verlag) pp 54-79
- Kapur D, Shyamasundar R K 1994 Software for safety-critical systems: Quality and futuristic technologies. Invited Lecture at the Seminar on Science Policy, Jawarharlal Nehru University, New Delhi
- Koymans R, Shyamasundar R K, de Roever W P, Gerth R, Arun Kumar S 1988 Compositional semantics for real-time distributed computing. *Inf. Comput.* 79: 210-256
- Ladeau B R, Freeman C 1991 Using formal specification for product development. *Hewlett-Packard J.* 6: 62-66
- Lamport L 1983 What good is temporal logic. *IFIP*
- Leveson N G 1986 Software safety: Why, what and how. *ACM Comput. Surv.* 18: 125-163

- Leveson N G 1991 Software safety in embedded computer systems. *Commun. ACM* 34: 34-46
- Mojdehbakhsh R, Tsai W T, Kirani S, Elliott L 1994 Retrofitting software safety in an implantable medical device. *IEEE Software* : 41-50
- Parnas D, van Schouwen J, Po Kwan S 1990 Evaluation of safety-critical software. *Commun. ACM* 33: 636-648
- Pnueli A 1986 Specification and development of reactive systems. *IFIP* : 845-858
- Pnueli A 1992 System specification and refinement in temporal logic. *Proc. FST & TCS 92, Lecture Notes in Computer Science* (Berlin: Springer-Verlag)
- PRG 1992 Programming Research Group at Oxford. *Timed CSP: Theory and practice, Lecture Notes in Computer Science* (Berlin: Springer-Verlag) 600: 640-675
- Place P R H, Wood W G, Tudball M 1990 Survey of formal specification techniques for reactive systems. CMU/SEI-90-TR-5, Software Engineering Institute, Carnegie Mellon University, Pittsburgh
- Rushby J 1993 Formal methods and the certification of critical systems. Computer Science Lab Tech. Report SRI-CSL-93-07, SRI Intl, Menlo Park, CA
- Shyamasundar R K 1993 Specification of hybrid systems in CRP. *Proc. of the 3rd Int. Conference on Algebraic Methodology and Software Technology (AMAST 93)* University of Twente, The Netherlands; full proceedings in the *Workshops in Computing Series* (eds) M Nivat, C Rattray, T Rus, G Scollo (Berlin: Springer-Verlag) pp 227-238
- Shyamasundar R K 1994 Specifying dynamic real-time systems in CRP. *IFIP 94 Congress Hamburg*
- Simon H 1969 *The sciences of the artificial* (Cambridge: MIT Press)
- Tanenbaum A 1981 *Computer networks* (New York: McGraw-Hill)
- Weber-Wulff D W 1993 Selling formal methods to industry. *Formal Methods Europe 93, Lecture Notes in Computer Science* (Berlin: Springer-Verlag) pp 671-678



# A survey of Indian logic from the point of view of computer science

V V S SARMA

Department of Computer Science & Automation,  
Indian Institute of Science, Bangalore 560012, India

**Abstract.** Indian logic has a long history. It somewhat covers the domains of two of the six schools (*darsanas*) of Indian philosophy, namely, *Nyaya* and *Vaisesika*. The generally accepted definition of Indian logic over the ages is the science which ascertains valid knowledge either by means of six senses or by means of the five members of the syllogism. In other words, perception and inference constitute the subject matter of logic. The science of logic evolved in India through three ages: the ancient, the medieval and the modern, spanning almost thirty centuries. Advances in Computer Science, in particular, in Artificial Intelligence have got researchers in these areas interested in the basic problems of language, logic and cognition in the past three decades. In the 1980s, Artificial Intelligence has evolved into knowledge-based and intelligent system design, and the knowledge base and inference engine have become standard subsystems of an intelligent system. One of the important issues in the design of such systems is knowledge acquisition from humans who are experts in a branch of learning (such as medicine or law) and transferring that knowledge to a computing system. The second important issue in such systems is the validation of the knowledge base of the system i.e. ensuring that the knowledge is complete and consistent. It is in this context that comparative study of Indian logic with recent theories of logic, language and knowledge engineering will help the computer scientist understand the deeper implications of the terms and concepts he is currently using and attempting to develop.

**Keywords.** Indian logic; logic; language; artificial intelligence; cognition.

## 1. History

*Mahamahopadhyaya* Satis Chandra *Vidyabhusana* in his monumental work on the history of Indian Logic considers three principal phases of developments: The ancient school of Indian logic with the representative text, *Nyaya Sutra* of Gautama (650 B.C. - 100 A.D.), the Medieval school of Indian logic with the representative

text, *Pramana Samuccaya* of Dignaga (100 A.D. - 1200 A.D.) and the modern school with the representative text, *Tattva Cintamani* of Gangesa (900 A.D. onwards). The Hindus and the Greeks have developed their logical systems largely independent of each other. It is conceivable that the notion of a syllogism developed by Aristotle in his Rhetoric might have found its way to India. It has been said that sage Narada visited Svetadvipa (Alexandria) and became an expert in the handling of the five limbed syllogism. (Chandra *Vidyabhusana*, 1921).

*Anviksiki* started as the science of inquiry and has grown into the art of debate. It had its beginnings in the *Atma Vidya* or *Brahma Vidya* (science of the soul or the divine science) pursued by the *Upanisads*. *Anviksiki* differed from *Atma Vidya* as it dealt with two subjects: *atma*, the soul and *hetu*, the theory of reasons. It later bifurcates into philosophy and logic. In the former aspect, it evolved into *darsana* and in the later aspect it evolved into *hetu vidya* (the science of reasoning) or *tarka vidya* (the art of debate). *Anviksiki* has been held in high esteem in works such as Kautilya's *Artha sastra*.

The technical terms of *anviksiki* may be found in books such as *Aitareya Brahmana*, and *Kathopanishad*. One can visualise a council conducting debates of learned men (*Samsad*, *samiti*, *sabha* or *parisad*), where discussions on true knowledge were taking place in the context of four valid means of obtaining the same:

1. *Smrti* (scripture);
2. *Pratyaksha* (perception);
3. *Aitihya* (tradition);
4. *Anumana* (inference).

## 2. The ancient school

Indian logic by itself is a very vast subject. The ancient Sanskrit term, *Nyaya*, probably was used initially in a more general sense. *Vaisesika* is often considered the sister science (*sastra*) of *Nyaya*. The purpose of Indian logic has been to acquire valid knowledge through perception and inference.

The *Nyaya* sutras are associated with Aksapada Gautama. The five books on *Nyaya* deal with the following:

1. Sixteen categories of the system;
2. Doubt, the four means of proof and their validity;
3. The self, the body, the senses and their objects, the mind and cognition;
4. Theory of error and of the whole and its parts; and
5. Unreal objections and occasions for rebuke of an opponent.

The *Vaisesika* works are traced to Kanada. These deal with, among other things,

1. The five categories: substance, quality, motion, generality and particularity;
2. The five elements: earth, water, fire, air, ether and space and time;

3. Objects of sense, the self, the mind and the theory of inference;
4. Perception;
5. Causality ;
6. The atomic theory, the self and inherence.

While sciences such as astronomy, geometry and philology have arisen in close connection with the sacrificial rituals of Vedas, Nyaya probably arose in the context of the *Mimamsa*. Anviksiki as the name of a science appears in *Gautama Dharma Sutra* beside the Vedic science (*trayi*). By means of this Vyasa is said to have arranged the Upanisads as recorded in the *Mahabharata*. Anviksiki has been treated as a subject of study suitable for a king. It was probably applied to secular ends such as justice apart from being applied to sacred things. It was probably the reason why Anviksiki was censured in the *Ramayana*, as its wrong application was leading men towards not following the prescriptions of the *Dharmacastras*.

Sage Narada is described as skilled in Nyaya, able to distinguish unity and plurality, conjunction and inference, priority and posteriority, deciding matters by means of proof, and a judge of the merits and demerits of a five-membered syllogism. Caraka in his medical *Samhita* gives a sketch of some of the Nyaya principles, and of the Vaisesika categories, in such a way as to indicate that he regarded the systems as supplementing each other. In the earlier grammatical literature, Panini, Katyayana and Patanjali know the meaning of Nyaya as conclusion but show no trace of recognizing a Nyaya system.

## 2.1 Knowledge

Aksapada Gautama says in his Nyaya Sutras that supreme felicity is obtained by the knowledge of the sixteen categories treated in his work:

### 1. The means of right knowledge (*pramana*)

Perception (*pratyaksa*), inference (*anumana*), comparison (*upamana*) and word or verbal testimony (*abda*) are the means of right knowledge.

- (a) Perception is the knowledge which arises from the contact of a sense with its object, being unnameable, determinate and nonerratic.

Un-nameable <i>avyapadesyam</i>	Signifies that the knowledge of a thing derived through perception has no connection with the name the thing bears
Deterministic <i>vyavasayathmakam</i>	This distinguishes perception from uncertain knowledge (from a distance a man cannot distinguish between dust and smoke)
Nonerratic <i>Avyabhicari</i>	Man is prone to visual illusions. In summer, one may see water far away in a mirage, when there is no water.

There is thus a need to establish the validity of the knowledge obtained through perception *Pramanya vada*. *Prama* stands for valid knowledge, while *aprama* or *bhrama* stand for invalid knowledge.

- (b) Inference is knowledge which is preceded by perception and can be a *priori*, a *posteriori* or commonly seen.

Gautama lays down that there are five members (*avayava*) of a syllogism, namely,

- i. proposition (*pratijna*),
- ii. reason (*hetu*),
- iii. example (*udaharana*),
- iv. application (*upanaya*),
- v. conclusion (*nigamana*).

The scheme of Gautama is illustrated by the syllogism:

The hill is fiery.

Because it has smoke.

Whatever is smoky is fiery, like a kitchen.

So is the hill (smoky).

Therefore, the hill is fiery.

- (c) Comparison is the knowledge of a thing through its similarity to another thing which is previously well known. This is often called "Learning by Analogy" in AI literature.
  - (d) Verbal Testimony or Word (*sabda*) is the instructive assertion of a reliable person. Learning from an expert (Knowledge Acquisition) in a modern day expert system like MYCIN (a computer program for diagnosing and prescribing treatment of infectious diseases) is often done by the knowledge engineer interviewing an expert and noting his situation - action behaviour, and representing it as an IF-THEN rule (Rich, 1985).
2. The object of right knowledge (*prameya*). Soul (*atma*); body (*sarira*); senses (*indriya*) : nose, tongue, eye, skin. ear; intellect (*buddhi*); mind (*manas*) and some others are listed under this.
  3. Doubt (*samsaya*) is a conflicting judgment about the precise character of an object.
  4. Purpose (*prayojana*).
  5. Example (*drstanta*).
  6. Tenet (*siddhanta*).
  7. Members (*avayava*).
  8. Confutation (*tarka*).
  9. Ascertainment (*nirnaya*).
  10. Discussion (*vada*) is the adoption by two parties of two opposite theses, which are each analysed in the form of the five-membered syllogism, and are supported or condemned by any of the means of right knowledge, and by confutation without deviation from the established tenets. The objective of discussion is seeking the truth.



Example:

Discutient (D): There is soul.

Opponent (O): There is no soul.

D: Soul is existent (proposition).

Because it is an abode of consciousness (reason).

Whatever is not existent is not an abode of consciousness, as a hare's horn (negative example).

Soul is not so, that is soul is an abode of consciousness (negative application).

Therefore, soul is existent (conclusion).

O: Soul is nonexistent (proposition).

Because it is not perceptible by any of the senses (reason).

Whatever is not perceptible by any of our senses is not existent as a hare's horn (positive example).

Soul is so (is not perceptible by any of our senses) (positive example).

Therefore soul is nonexistent (conclusion).

D: The scripture which is a means of right knowledge declares the existence of soul.

O: The scriptures (of certain sects) deny the existence of soul.

11. Wrangling (*jalpa*) aims at gaining victory by resorting to quibbles, analogues and processes which deserve rebuke.
12. Cavil (*vitanda*) consists of mere attacks on the arguments of the opponent.
13. Fallacies (*hetvabhasa*) of reason are the erratic (*savyabhicara*), the contradictory (*viruddha*), the controversial (*prakarana sama*), the counter questioned (*sadhya sama*) and the mistimed (*kalatita*).

Examples:

Sound is noneternal (proposition). Because it is not possessed of the attribute of eternality (reason).

The reason does not throw any new light.

14. Quibble (*chala*).
15. Analogue (*jati*).
16. The point of defeat (*nigrahasthana*).

Vatsyayana, author of *Nyaya Bhashya*, one of the many commentaries of the Nyaya Sutra (A.D. 400), reveals that there were others who raised the number of members of the syllogism to ten. They are the desire to know (*jijnasa*), the doubt (*samcaya*), the belief in the possibility of a solution (*cakyaaprapti*), the purpose in view in attaining the solution (*prayojana*), the removal of doubt (*samcaya-vyudasa*).

With its full ten members, we have before us in miniature, the course of the kind of discussions which preceded the development of the logical process, and we can recognize the substantial progress achieved in omitting all that did not directly bear on the attainment of a conclusion.

Note that a modern Artificial Intelligence text typically talks briefly only of Aristotle's three-membered syllogism involving logical deduction (*modus ponens*) (Rich 1985).

Example:

All men are mortal.

Socrates is a man.

Therefore, Socrates is mortal.

While deduction is a perfectly valid proof, real life systems do not permit its use excepting in applications such as mathematical theorem proving. In a real life system such as medical diagnosis or pattern recognition, one has to be content with reasoning techniques such as induction and abduction. Induction involves generalisation from examples and abduction provides a plausible explanation.

### 3. The medieval school

Buddhist philosophers Nagarjuna (250-300 A.D.) and Maitreya (400 A.D.) use three-membered syllogisms:

1. The hill is full of fire.
2. Because it is full of smoke.
3. That which is full of smoke is full of fire, as a kitchen.

Vasubandhu (about 450 A.D.) omits the example and gives his syllogism as follows:

1. The hill is full of fire.
2. Because it is full of smoke.
3. All that is full of smoke being full of fire.

Buddhist logician Dignaga (450-520 A.D.) is considered to be the greatest logician India has ever produced. He was born in Kancheepuram, lived in *Vengi desa* (present West Godavari District, A.P.) and later probably travelled north. He had several works to his credit such as *Pramana Samuccaya*, *Nyaya Pravesa*, and *Hetu Cakra Samarthana*. Most of the works of Buddhist logicians of this age are available only in Tibetan.

*Pramana Samuccaya*, begins thus:

Bowing down before Sugata, the teacher and protector of the world, I, for the sake of expounding valid knowledge (pramana), put together here various scattered matters, compiled from my own works. The book has six chapters entitled

1. Perception (*pratyaksa*).
2. Inference for own self (*svarthanumana*).

3. Inference for the sake of others (*pararthanumana*),
4. Reason and Example (*hetu, drstanta*),
5. Negation of opposite (*apoha*),
6. Analogue (*jati*).

Dignaga notes that demonstration and refutation, together with their fallacies, are useful in arguing with others while perception and inference together with their fallacies are useful for self understanding.

S	P	R	E1	E2
Subject	Predicate	Reason or mark	Example 1	Example 2
A minor term <i>paksa</i>	A major term <i>sadhya</i>	A middle term <i>hetu or linga</i>	<i>drstanta</i> homoge- neous	<i>drstanta</i> hetero- geneous
or <i>dharmin</i>	or <i>dharma</i>	or <i>sadhana</i>	<i>sadharmya</i>	<i>vaidharmya</i>

The form of syllogism is as follows:

1. This hill (S) is fiery (P).
2. Because it has smoke (R).
3. All that has smoke is fiery, like a kitchen ( *homogeneous example*, E1) and whatever is not fiery has no smoke, like a lake( *heterogeneous example*, E2).

A proposition offered for proof is a thesis. The following are some of the fallacies of the thesis.

1. A thesis incompatible with perception. (e.g. Sound is inaudible.)
2. A thesis incompatible with inference. (e.g. A pot is eternal. In reality, it is noneternal because it is a product.)
3. A thesis incompatible with public opinion. (e.g. Money is an abominable thing. While some saints may hold this opinion, the world does not say so. In fact, it says *Dhana mulam idam jagat*.)
4. A thesis incompatible with one's own belief or doctrine. (e.g. A Vaisesika philosopher saying that sound is eternal.)
5. A thesis incompatible with one's own statement. (eg. My mother is barren.)

Another great logician of this school is Dharmakirti (around 650 A.D.). His works include *Pramana Vartika Karika and Nyaya Bindu*. His definition of perception as a source of valid knowledge is an improvement over the ones in the earlier age. While valid knowledge can be acquired through senses, Dharmakirti says, it should be free from preconception (*kalpana*) and devoid of error (*abhanta*). For example, in darkness a rope might appear as a snake, and for a person moving in a boat, trees on the bank may appear to be moving in the opposite direction. In other words, these logicians were very much concerned with uncertainty in human knowledge, which incidentally is a major research issue in modern day knowledge system design. Dharmakirti also discusses the requirements of the middle term smoke in the context of *svarthanumana*.

1. The hill has fire.
2. Because it has smoke.
3. Like a kitchen, but unlike a lake.

### 3.1 Jaina logic

Apart from Buddhist scholars, Jaina scholars also have contributed significantly to logic in this period. In the Jaina scriptures, *Sthananga-sutra* and *Bhagavati-sutra*, there is classification of valid knowledge as *pramana*, *pamana*, *jnana*, *nana* or *hetu*. When *hetu* is used in the sense of inference, it is classified according to the following types:

1. This *is* because that *is* : There is a fire, because there is smoke.
2. This *is not* because that *is* : It is not cold, because there is a fire.
3. This *is* because that *is not* : It is cold here, because there is no fire.
4. This *is not* because that *is not* : There is no *simsupa* tree here, because there are no trees at all.

Bhadrabahu, in his *Sutra-krtanga nirvyukti*, mentions another principle of the Jaina logic called *Syadvada*, or the assertion of possibilities. The *Syadvada* is set forth as follows:

1. May be, it is.
2. May be, it is not.
3. May be it is, and it is not.
4. May be it is indescribable.
5. May be it is, and yet is indescribable.
6. May be it is not, and yet is indescribable.
7. May be, it is and it is not and it is indescribable.

## 4. The modern age of Indian logic

Gangesa Upadhyaya of Mithila of 12th century is a key figure representing the modern Indian school of logic. He is the author of *Tattva Cintamani*, which consists of four books dealing respectively with perception (*pratyaksa*), inference (*anumana*), comparison (*upamana*) and verbal testimony (*sabda*), which are the four means of obtaining valid knowledge.

### 4.1 Perception

Gangesa in the book I of *Tattva cintamani* distinguishes between ordinary perception (*laukika pratyaksa*) and transcendental (*alaukika pratyaksa*). The latter, in turn, may be having *samanya laksana*, *jnana laksana* or *yogaja laksana* corresponding to ordinary or enlightened (or transcendental) characteristics.

The ordinary perception is of six kinds:

1. Union (*samyoga*): In the visual perception of a jar or chair, there is a union of the eye of the observer with the object called a jar or a chair.
2. United-inherence (*samyukta-samavaya*): When we see a jar, we also see its colour, which is an inherent attribute of the jar.
3. United-inherent-inherence (*samyukta-samaveta-samavaya*): When we see jar, we see its colour and also the concept or notion of an object having a colour.
4. Inherence (*ṣamavaya*): Sound is inherently perceived by our ears.
5. Inherent-inherence (*samavetha-samavaya*): The soundness of sound is also perceived.
6. Particularity (*visesanata*): Perception of nonexistence of a jar when a jar is not there.

The sense through the instrumentality of which we perceive colour is the eye. Similarly, we perceive sound with the ear and smell with the nose. These are examples of external senses. The sense which operates as an instrument, in our perception of pleasure, pain, desire, averice, intellect and volition is the MIND, which is called the internal sense. It is called atomic, since it can perceive objects one at a time. (e.g. The object is perceived either as a snake or a rope at a time.)

One can also distinguish between immediate perception and mediate or reflective perception.

Example:

This is a pot.

I know this is a pot.

I know this object is a pot as I can perceive its potness.

### 4.2 Inference

The second chapter of Gangesa's book deals with inference (*anumana khanda*).

Gangesa agrees that inference is one of the means of generating knowledge (*anumiti-nirupana*). The most interesting contribution in his work on inference is the doctrine of invariable concomitance (*vyapti*). It has been described variously as pervasion, inseparable connection, perpetual attendance or constant copresence.

There are five provisional definitions for this doctrine (*vyapti panchakam*). We shall see one of them.

## Definition

Invariable concomitance is the nonpresence of the middle term in the locus of the nonexistence of the major term.

If we consider the statement

The hill is full of fire because it is full of smoke. In this smoke is the middle term (R) and fire (P) is the major term.

$\{\text{fire, smoke}\} \subset \{\text{fire}\} \subset \{\text{no fire, no smoke}\}$

### 4.3 Navya-Nyaya

Though Nyaya and Vaisesika are separate systems, they have more similarities than dissimilarities. The later Nyaya school called, *Navya-Nyaya* (NN) developed as a result of blending of the two.

As an example, let us examine the following from the work of Viswanatha Nyaya-Pancanana (1634 A.D.). He belonged to the *Navadvipa* School of Bengal. He is credited with the Vaisesika treatise, *Bhasa-Pariccheda*, and the book on logic, *Siddhanta Muktaavali*.

Matter (*Padartha*) has seven categories: substance (*dravya*), quality (*guna*), action (*karma*), generality (*samanya*), particularity (*visesa*), inherence (*samavaya*), and nonexistence (*abhava*). This belong to *visaya kanda* of Vaisesika system.

Substance is composed of five elements, earth (*ksiti*), water (*apa*), light (*tejas*), air (*marut*), and ether (*vyoma*), space (*dik*), time (*kala*), mind (*manas*) and soul (*atma*).

Soul (*atma*) possesses intellect (*buddhi*) which comprises of apprehension or understanding (*anubhuti*) and remembrance or memory (*smrti*).

Understanding is due to perception (*pratyaksa*), inference (*anumana*), comparison (*upamana*) and verbal testimony (*śabda*). This belongs to *Jnana kanda* of Nyaya.

The *Tarka-samgraha* of Annambhatta is the most popular introductory work on the Nyaya-Vaisesika system of Indian philosophy. A native of Andhra, he flourished in 17th century (Virupakshananda 1980).

Mullatti (1977) represents the Navya-Nyaya theory of inference in terms of contemporary logic framework. He observes that the NN theory is couched in terms of cognitions rather than premises. So the theory demands adequate sentences (*pramana vakya*). These must satisfy four criteria: expectancy (*akanksa*), competency (*yogyata*), proximity (*sannidhi*) and speaker's intention (*tatparya*). In the sentence, "Bring a cow", the use of bring is said to raise an expectancy in the listener. A sentence such as "Bring triangularity" does not fulfil this requirement. The stock examples in the case of competency are:

1. *jaleṇa sincati* "(He) wets (the ground) with water."
2. *agnina sincati* "(He) wets (the ground) with fire."

While both are syntactically sound, only the first is semantically sound. It is sufficient, if the sentence is sound. It need not be true. The NN theory does not also accept unexampled terms such as "barren woman's son" (*vandhya-suta*), "hare's horn" (*sasa-srnga*), and "sky-flower" (*gagana-kusuma*). Proximity refers to the ambiguity caused by word order. For example, observe the English sentences.

1. I saw a girl in the park with a telescope.

2. I saw a girl in the park with a dog.
3. I saw a girl in the park with a statue.

Probably these mean respectively,

1. *I saw with a telescope* a girl in the park.
2. I saw *a girl with a dog* in the park.
3. I saw a girl in *the park with a statue*.

The intention (*tatparya*) comes into play when we talk of a word like "door". The word "close" or "open" should be supplied to find the intention of the speaker.

## 5. Truth

Indian philosophers maintain that there are truths beyond man's normal experience. Neither sense perception nor inference can impart knowledge of suprasensible facts. e.g. God, soul, their relation, soul's journey after death, heaven, hell, merit and demerit accruing from righteous and unrighteous deeds. How does one go beyond the truths that are beyond the range of the senses and reasoning?

In Vedic view, the cosmic order is controlled by a fundamental principle or truth called *Rta*. As stated in the *Rg-Veda*, the whole universe is founded on *Rta* and moves in it. Because of *Rta*, fire burns, wind blows, water flows, plants grow, humans beings think and seasons revolve. Untruth is *unrta*. The word *satya* for truth is often used in the restricted context of right speech.

The validity of the Vedic testimony is due to the fact that it discloses truths which can neither be contradicted nor established by any other means (Satprakashananda 1965).

Karl Popper's dictum clarifies the difference between truth and certainty: "We must distinguish between TRUTH, which is objective and absolute and CERTAINTY, which is subjective."

All elements of a knowledge base are thus uncertain to a more or less degree as these refer to chunks of human expert knowledge. On the other hand in logic, one is traditionally concerned with truth. In a propositional logic system, one should be able to identify a proposition as TRUE or FALSE. Examples of propositions are:

1. Panini is a grammarian.
2. Annambhatta is the author of *Tarka Samgraha*.
3. If  $2 + 2 = 6$ , then I am prime minister of India.
4. The present king of India is bald.
5. Rama is tall.

While it is easy to associate truth values to propositions 1 and 2 above, the truth of a compound proposition like 3 is to be derived while proposition 4 poses more problems. Truth itself may be a matter of degree, as in proposition 5. Truth may often have to be established by combining unreliable evidences given by multiple

sources. Logic as a tool of knowledge representation and reasoning and handling of uncertainties in knowledge are topics currently researched actively in Artificial Intelligence.

To illustrate the difference between truth and belief let us present the experience of Brancazio (1994):

A Moslem student noted that Quran clearly states that Jesus Christ was not crucified. This, of course, provided considerable protest from Christian students. They were surprised to learn that outside of the gospels, there is no independent secular historical record on the trial or crucifixion of Jesus. What was one to conclude? Or, to state the question dramatically, what is truth?

It is necessary to talk about two theories of the nature of truth. The correspondence theory claims that a statement is true if it corresponds to objective reality. The coherence theory holds that a statement is true if it is consistent with other true statements in a self-consistent system of ideas and concepts. It can be argued that the correspondence theory of truth is the model used by science (physical sciences) whereas the religious truth follows the coherence theory in the sense that a statement is true if it is consistent with the belief system of a religious community. It is interesting to note that the coherence theory of truth is the model used by mathematicians and logicians.

For the author who is neither a Moslem nor a Christian, it is sensible to assume that Jesus was either crucified or was not with a certainty factor 0.5 based on equal positive and negative evidence. Others might fall back on the coherence theory and state that crucifixion is true in the Christian world while it is not true in the Moslem world.

## 6. Knowledge representation and Sanskrit

Computer processing of natural languages, as opposed to artificial languages such as Fortran, Pascal and C used for writing computer programs, is an active area in Artificial Intelligence. It is now understood that this is an extremely difficult task. It has now been well recognized that Panini's Sanskrit grammar *Astadhyayi* presents a framework for a universal grammar of any language. This rule and meta rule based grammar uses ideas of recursion almost twenty centuries before the idea of a computer program. Its equivalence to the powerful knowledge representation structures such as semantic nets have been recognized in recent times (Briggs 1985; Kak 1987). In particular, the similarities between the *karaka* theory of Panini and the conceptual dependency and conceptual graph approaches used in Artificial Intelligence may be noted (Rich 1985).

Panini took the idea of action as described by a verb and developed the *karaka* theory by providing a context for action in terms of its relations to agent and situation.

<i>Apadana</i>	That which is fixed when departure takes place
<i>Sampradana</i>	The recipient of the object
<i>Karana</i>	The main cause of the effect, instrument
<i>Adhikarana</i>	The basis, location
<i>Karman</i>	The object
<i>Kartr</i>	The agent, the independent actor



## 7. Concluding remarks

In this paper, we attempted to give a flavour of classical Indian logic as it evolved over more than thirty centuries. It is interesting to know the context in which ancients viewed issues such as truth, knowledge, intelligence, cognition and language. The engineering community, of late, has started learning these concepts while attempting to design intelligent systems possessing a knowledge-base acquired from domain experts, together with the uncertainties associated with such knowledge and an inference engine performing automated reasoning. The Nyaya-Vaisesika combination is analogous to such an exercise. While most of modern research in Artificial Intelligence owes its origin to cold war and military applications, the nobler aims of the ancient scholars in acquiring knowledge may also be examined by present day scholars.

## References

- Brancazio P J 1994 What is truth? A course in science and religion. *Am. J. Phys.* 62: 893 - 898
- Briggs R 1985 Knowledge representation in Sanskrit and Artificial Intelligence. *AI Mag.* 6: 22-38
- Chandra S *Vidyabhusana* 1970 *A history of Indian logic* (Delhi: Motilal Banarsidass) (First published by Calcutta University in 1921)
- Kak S C 1987 The Paninian approach to natural language processing. *Int. J. Approximate Reasoning* 1: 117-130
- Mullatti L C 1977 *The Navya-Nyaya theory of inference* (Dharwad: Karnatak University)
- Rich E 1983 *Artificial intelligence* (New York: McGraw-Hill)
- Satprakashananda Swami 1965 *Methods of Knowledge according to Advaita Vedanta* (London: George Allen & Unwin)
- Virupakshananda Swami 1980 *Tarkasamgraha (with Dipika of Annambhatta)* (Madras: Sri Ramakrishna Math)



# Rudiments of complexity theory for scientists and engineers

V VINAY

Department of Computer Science and Automation,  
Indian Institute of Science, Bangalore 560 012, India

**Abstract.** Complexity theory is an important and growing area in computer science that has caught the imagination of many researchers in mathematics, physics and biology. In order to reach out to a large section of scientists and engineers, the paper introduces elementary concepts in complexity theory in a informal manner, motivating the reader with many examples.

**Keywords.** Complexity theory; computational complexity; algorithms.

## 1. Introduction

All of us know how to calculate the *determinant* of a  $4 \times 4$  matrix. We would use the standard row expansion with alternating signs. Using this algorithm involves computing the determinant of four  $3 \times 3$  matrices or  $12 \times 2 \times 2$  matrices. The number of  $2 \times 2$  matrices we need to compute becomes rather large as we increase the dimension of the matrix. In fact, an  $n$ -dimensional determinant would need  $n!/2$  evaluations. However, we are also familiar with an easier way to solve the problem: use elementary row and column transformations to convert the given matrix into an upper triangular matrix and then simply multiply the diagonal entries. It can be argued that effort in terms of the number of multiplications and divisions involved is no more than  $n^3$  operations. Though we know two ways to solve the same problem, we intuitively know that the latter method is preferable for large-dimensional matrices. Because the *complexity* in the number of steps is smaller.

Consider now a slightly different problem, the problem of computing a *permanent* of the same matrix. Its definition is similar to that of the determinant but with one important change: ignore the alternate sign change convention. Instead, all signs are taken to be positive. The row expansion method solves the problem but is again costly. Unfortunately, the elementary row and column transformations do not preserve the permanent of the matrix. We are now stuck with a very time-consuming method. Is there a better, more efficient way of calculating the permanent? Nobody

---

\*Part of the work was done while the author was with the Centre for Artificial Intelligence and Robotics, Bangalore

knows of any substantially more efficient method for doing so. Probably there are none! Can such statements be made with our current knowledge?

These are the nature of questions we want to address in this article. The article can be read by anybody who is exposed to elements of programming and who has used a computer. All concepts are introduced through examples, as far as possible. There are very few book on the subject. Readers who want to pursue their reading further can refer to Garey & Johnson (1979), Cook (1985) and Balcazar *et al* (1990).

## 2. Where does complexity theory stand?

The role of any scientific theory is to interpret and predict phenomena within its realm. Scientific theories act on a well-defined domain within which it interprets and predicts. A botanist finds a natural domain in flora and fauna. Mathematics creates abstract domains to investigate. Thus, for example, space need not be Euclidean. The domain may, of course, itself undergo refinements over time to accommodate for a deeper understanding. Thus, the domain of Newtonian mechanics is different from the domain of quantum mechanics, in that the latter refines the former in dramatic ways. In short, natural sciences seem to have "God given" domains whereas mathematics seems to create its own domain as and when required. A criteria often used is that natural sciences have observable domains. I wish to put forth the view that *computing is a natural one*. Unlike in mathematics, it is not created by axioms. And unlike in natural sciences, it does not have a physically observable domain. But the domain of all problems, is a "God given" one; and to that extent *the science of computing is a natural science*. And to the extent that the domain exists in the abstract, *it is mathematics*. Either way, the *science of computing* is a science, worthy of attention and contemplation by the best of minds.

Depending on one's perspective, *complexity theory* either subsumes or is subsumed by the science of computing. Complexity exists in every field. For example, just as the structural properties of the determinant was useful in designing an efficient method to evaluate the determinant, structural properties of chemicals can be used to speed the chemical process by using (for example) a catalyst. For the purposes of this paper, we shall restrict ourselves to the world of computing.

It should be emphasized that the complexity of solving a problem is *inherent* to the problem *per se* and does not depend on the nature of the computer we run the algorithm on, the programming language we use, and so on. The science of computing exists irrespective of the notion of a computer just as the notion of time is independent of a clock. Computers and clocks are realizations that give form to the abstractions involved.

## 3. Questions in complexity theory

One of the main spin-offs of any reasonable theory is that it provides a framework to classify or group like objects. Typically, the classification depends on a small set of parameters. Complexity theory is no different. First, a model of computation is fixed. Resources are identified in the model of computation. By varying the resources, different *complexity classes* result. Essentially, a complexity class (associated with a model of computation) is the set of all problems that can be solved on the model within the stated resource. Not all models of computation have physical

realizations. However these models are important because they capture the essence of the complexity of the problem. We will quantify the sentence in a later section.

We will now list some of the main questions that require to be addressed. They are,

- How fast can a problem be solved?
- How fast can a problem be solved with the given resources?
- How do resources define complexity classes?
- Is the complexity class so defined robust?
- Is the problem intrinsically difficult to solve? If so, in what way?
- What properties does a complexity class satisfy?
- What can be said about relationships among complexity classes?

The first two questions are addressed by *design and analysis of algorithms*. The other questions come under *complexity theory*.

We have to explain some of the terms used above. When can we regard a complexity class as robust? As was explained above, a complexity class is the set of all problems that can be solved on an associated model of computation within some resource bounds. It is possible that a different model of computation with appropriate resource bounds on it also defines the same set of problems. In a happy situation such as this, we realize that the problems in the set that form a complexity class is “model independent”, i.e., the intrinsic property (or properties) shared by the problems in the set is *not* specific to a model of computation. We can now say that a complexity class has alternate characterizations. Most complexity classes are robust. In fact, alternate characterizations of some complexity classes are positively startling and totally unexpected.

Alternate characterizations are also useful as they provide new insights into the problem and its complexity. These insights are valuable in identifying key structural properties of the problem in hand.

What do we mean by a problem being difficult to solve? In order to go any further, we need some definitions.

We imagine all our problems are encoded in binary. A problem instance is then a binary string. For technical reasons, we confine ourselves to the so-called *decision problems*: problem instances whose outcome is a “yes” or a “no”. The “yes” instances of a problem constitute a set  $L$ . We will refer to  $L$  as either a problem or a language. In many complexity class, a notion of *completeness* can be introduced. A problem is said to be complete for a complexity class if

1. it belongs to the complexity class, and
2. it is as *hard* as any other problem in the class.

If  $C$  is a complexity class and  $L$  is said to be  $C$ -complete if  $L$  satisfies the two properties listed above. If, however, only property (2) is satisfied,  $L$  is said to be  $C$ -hard. The notion of being as hard as any other problem is technical – I will only briefly sketch what it means. Let  $L_1$  and  $L_2$  be two languages. We say  $L_1$  is

*reducible to  $L_2$*  if there is a mapping wherein every string in  $L_1$  is mapped to some string in  $L_2$  and every string that is not in  $L_1$  is mapped on to some string not in  $L_2$ . A problem is hard for a class if every problem (or language) in the class is reducible to the given problem. The mapping that witnesses the reducibility should be "simple" or "easy".

Suppose  $L_1$  and  $L_2$  are  $\mathcal{C}$ -complete. Then, by definition,  $L_1$  and  $L_2$  are the hardest problems in  $\mathcal{C}$ ; and so  $L_1$  is as hard as  $L_2$  and vice versa. We may conclude that the two problems have the same "degree" of hardness.

All this is important as it allows us to pick *one* representative problem from a complexity class and investigating its properties gives us insight into the complexity class. This also allows one to refine the definitions of the complexity class, resulting in a deeper understanding.

Not all complexity classes seem to admit of a complete problem. Fortunately, many interesting complexity classes do. It makes things that much easier to investigate the class.

#### 4. Five examples

We will look at five examples. These problems have been chosen for simplicity. And each of them illustrate an interesting point.

1. Consider an  $m \times n$  grid. Our goal is to count the number of paths from a point  $s$  (lower left corner) to a  $t$  (upper right corner). The only restriction is that as we move from  $s$  to  $t$ , we may do so by either moving right or moving up. Therefore all paths have length  $m + n$ .

We want to count the number of paths from  $s$  to  $t$ . Elementary counting tells us that there are  ${}^{(m+n)}C_m$  paths, or  $(m+n)!/(m!n!)$ . Let us now provide a slight twist to the problem. Let us cut off some of the edges in the grid, as illustrated in figure 1. How do we now calculate the number of paths from  $s$  to  $t$ ?

If we were to patiently enumerate all the paths from  $s$  to  $t$ , we could then count the number of paths we enumerated. But we need to be really patient to make this method give an answer. To see this, consider a square grid of size 60. Let us cut a small number of connections in the grid. By removing only a few edges, we ensure that the number of paths in the present grid is nearly equal to the number of paths on the complete grid. On the complete grid, we know that the number of paths is  ${}^{60}C_{30}$  which is about  $10^{17}$ . Even assuming that a new path can be enumerated every nanosecond, the task would take about 3 years! A similar task on a  $100 \times 100$  grid would take almost forever.

The crux of the matter is that the number of paths grows *exponentially* in the dimension of the grid. (Assuming a square grid of size  $n$ , it is easy to see that  $2^n C_n \geq 2^n$ .) All this makes the method a very costly way of solving the problem. We require to find a much simpler way. Luckily, there is such a method. Consider any grid point. We can reach this grid point in exactly two ways: through its left neighbour or its neighbour below. (It is possible that both the neighbours do not exist, but at least one will.) Assume that we know the number of paths from  $s$  to each of the two neighbours is known. Clearly, the number of points to the given point is the sum of the number of paths to its two neighbours. If a neighbour does not exist, it simply means we treat it as a neighbour with count zero. The number of paths from  $s$  to itself is 1. The interested reader can work out the details and try

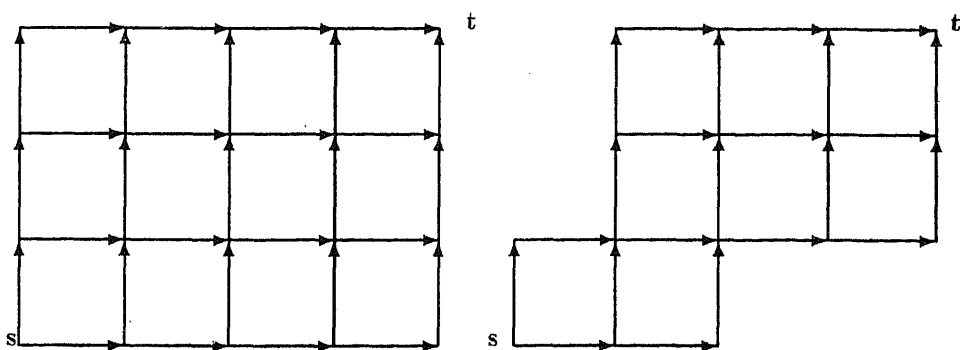


Figure 1. Counting paths between two nodes in a grid.

it out on a small example.

The number of additions to be carried out is no more than the number of grid points. This is because the moment we know the count of the neighbours of a point, we can determine the value of the point—so we need to look at each point exactly once. But a square grid of size  $n$  has  $n^2$  points. If we now assume that each addition takes one nanosecond, we see that the number of paths on some 100-dimensional grid is no more than  $10^{-5}$  seconds!

The reason for this improvement is because unlike the first method which takes *exponential time*, the second method takes *polynomial time*. Exponentials grow very fast and consequently many problems are almost impossible to solve, irrespective of the speed of the computer.

In an intuitive way, we can say a problem is *tractable* if it is solvable in polynomial time. That is, the graph plotting the size of the input to its time can be bounded above by a polynomial function.

2. Consider the problem of solving a jig-saw puzzle. All of us have solved one sometime or the other. In fact, it is common to find jig-saw puzzles with over a thousand pieces in the market. And they take hours to assemble. We want to find the complexity of solving this problem. All of us know the problem is not easy. The corners are easy to identify but in order to find pieces that go in the interior, we can do no better than trying out all possibilities until we find a piece that fits. This method is expensive and exhaustive. (You can easily convince yourself that there is an exponential effort involved.) Unfortunately, we do not know of any way to substantially speed up the time it takes to solve the puzzle.

One reason for this situation is that the problem appears to be lacking in structure that can help in finding a speedy solution. But it does have a property, a property that pleases and assures a complexity theorist—if somebody claims to have found a solution, you can easily verify the claim! I know this looks trivial, but hidden in it is one of the most profound ideas of our times.

The key is in the length of the solution as a function of the length of the input. In this case, the length of the input is the number of pieces. The length of the solution is linear (and hence polynomial) in the number of pieces. Let us write down two properties that the jig-saw puzzle has:

1. it has an easily describable solution;
2. the correctness of the solution is easily verifiable.

The word easy is used in the sense that the functions are polynomials. To us, polynomials are easy, tractable, efficient etc. In this example, we could not come up with an efficient method to solve the problem. But we identified two properties that the jig-saw puzzle satisfied. We replaced the criterion of "easily solvable" to "easily verifiable".

(3) Consider a solution to the previous example. Suppose it were written on a long sheet of paper and you are allowed to read it one word at a time (You can only remember a small number of words at any time). In the process of verifying the correctness of the solution, would you have to go back and forth over the written text? Let us see. Suppose there is a piece with three other pieces adjoining it. You verify that these are the three pieces that completely fit the fixed piece. Now you have to explore each of the three pieces further for correctness. While verifying the correctness of one of the three pieces, you would have to go back to the fixed piece as the fixed piece is a neighbour. You may argue that there is no need to check this up as it has already been verified. First of all, you cannot remember all the pairs you have checked as you have a small memory. (Before somebody jumps at me, I should state that I assume that each jig-saw has a constant number of neighbours; without this assumption, you may not even be in a position to verify the correctness of one fixed piece with respect to its neighbours!) Also, a moment's thought will convince you that this will allow your friend (who claimed the solution) to cheat! So it looks like we require to go back and forth on the presented solution. We will now consider an example where this need not be so.

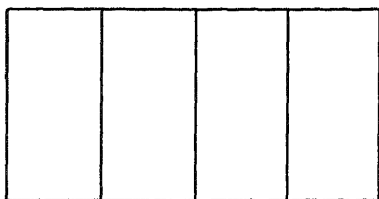
The city of Bangalore has suddenly become a motorists' nightmare, especially after half of its major roads were made one-way. Some motorists are even scared of going to some localities because they may not be able to figure out a way to get out! (In the next example, we will show that their fears are not unfounded.) In such a case, it is advisable to carry a route they need to take to reach their destination, at least on their first visit. The instruction written on a piece of paper contains descriptions of the following form: "As you go down the road, take the third right" or "At a circle, take the fifth road in clockwise direction". It is important to note that the motorist need never go back to a previous instruction; (S)he has to only remember the current instruction. The solution in this case is rather appropriately said to be *one-way*.

Problems with one-way solutions are far more easier than problems with "back and forth" solutions. In particular, such problems can also be solved in polynomial time. Problems with "back and forth" solutions are not known to be tractable (i.e., have polynomial time algorithms). This is one of the biggest open problems in complexity theory.

(4) We will consider the effects of making roads one-way. You as a motorist move from one road to another in a random fashion in a city without one-way roads. By random, we mean the following: if you are at an intersection and say six roads meet there, you will take a road with probability  $1/6$ . You will eventually reach your destination, of course. By an interesting application of probability theory, it can be proved that the number of steps (assuming one unit of time to go from one intersection to another) it takes on an average to reach the destination is a



polynomial in the number of intersections (or vertices in the graph). The proof is based on a simple analysis of a Markov chain.



We will now show that the number of steps need not be small in the presence of one-way roads. Consider a city with intersection and roads as represented by the graph above. Suppose we take a random walk starting at 0 and want to reach 9. A simple calculation shows that the probability of doing so is  $1/8$ . In a graph with  $n + 1$  vertices in each row, the probability works out to be  $1/2^n$ . So the expected number of steps required to reach the destination is exponential! This justifies the remark that was made in the previous section.

This example shows that randomness is also an useful computational resource. It can also be used to model the complexity of certain problems. The next problem shows another problem where randomization helps.

(5) Are the following two polynomials identical?

- $(3x^2 + 5x + 2)(4x^2 + 17x + 15)$
- $(x^2 + 4x + 3)(12x^2 + 23x + 10)$

If you expand each of the polynomials, you will require 18 multiplications and many additions to add like terms. In general, if the polynomial was a product of  $n$  arbitrary binomials, multiplying naively will result in  $2^n$  terms. One can be smarter and regroup like terms at each intermediate stage. But it is still a lot of work. We would very much like to have a simpler scheme. Assume the degree of the polynomial is  $n$ . Fix a large number, say  $2^n$ . Pick randomly a number between 0 and  $2^n$ . Evaluate the two polynomials at this random number. If the answers match, declare the polynomials as identical. Otherwise declare them as different.

When the evaluated answers are different, the polynomials are obviously different. However, it is possible that two different polynomials give the same answer at the point we choose. But no two distinct polynomials can agree on more than  $n$  points. Here is where randomness comes in. The probability that we pick one of these  $n$  bad points is  $n/2^n$ -a really small probability! This example illustrates how randomization helps in simplifying algorithms. Randomness is indeed a useful resource.

## 5. A simple model of computation

A *Turing machine* is a standard model of computation. I will not give a formal definition of a Turing machine, but verbally describe it.

A Turing machine

- has a mechanism to access the given input. Through this mechanism, we can read any any bit of the input. The input cannot be tampered with.

- has a blackboard to do its calculations. The blackboard can hold a finite amount of information at any time. The size of the blackboard is the number of bits it can hold.
- a set of very simple rules (the program) which determine the input bit to be read, the part of the blackboard to be read, and the changes that need to be made.
- whistles<sup>1</sup> when the computation is complete and it accepts the input.
- whistles differently when the computation is complete and it rejects the input.

Time and space are two common resources defined on a Turing machine. By one unit of *time* we mean executing one step of the program. So time taken by a Turing machine is the total number of steps it takes before it whistles. By *space* we mean the maximum number of bits written on the blackboard. This determines the size of the blackboard we require for a particular problem. (Note that the size may vary with the length of the input.)

We would like to look at other models of computation. Some of these models are attractive as it is easier to reason with them. Specifically, boolean circuits provide a rich framework to model computation. A boolean circuit is composed of many gates— typically OR gates, AND gates and NOT gates. These gates are interconnected among themselves in some manner. There are also some gates called the INPUT gates which feed inputs to some gates in the circuit. The circuit has a distinguished gate called the output gate. The value of the output gate determines whether the input is accepted or rejected. By convention, we assume that NOT gates are at the input level, i.e, the input to a NOT gate is a circuit input. One constraint that the interconnections should follow is that there cannot be cycles or feedbacks. Basically, this means that computation on a circuit has a sense of direction.

Two common resources defined on circuits are size and depth. The *size* of a circuit is the number of gates in the circuit. The *depth* of a circuit is the longest path among all paths from the output gate to any input gate. Unless otherwise stated, we will assume that the gates of the circuit have unbounded fan-in. By a *semi-unbounded circuit* we mean a circuit where-in the OR gates have unbounded fan-in and the AND gates have restricted fan-in. The restriction on fan-in depends on the complexity class we want to characterize. The notion of semi-unboundedness introduces an asymmetry in the roles played by the gates in the circuit.

A single circuit has a fixed number of inputs, whereas problem instances have varying length. Therefore we require circuit families - one circuit for inputs of each length.

(1) The class  $\mathcal{P}$  is the class of all tractable problems. Problems in this set are problems that can be solved in polynomial time on a Turing machine. Equivalently, they correspond to polynomial-sized circuit families. Important problems in  $\mathcal{P}$  are Linear programming, Network flow, Horn clause satisfiability. All these problems are complete for  $\mathcal{P}$ .

(2) The class  $\mathcal{DLOG}$  is the class of all problems that can be solved in  $\log n$  space ( $n$  is the length of the input). This means that the size of the blackboard is restricted to be logarithmic in the input length. Since only a logarithmic number of bits can be

---

<sup>1</sup> A multimedia Turing machine!

written on the board, there are only a polynomial number of distinct messages that can be written on the board. This fact can be used to see that  $DLOG$  is a subset of  $\mathcal{P}$ . An important problem in this class is the evaluation of arithmetic expressions.

(3) The class  $\mathcal{NP}$  is the class of all problems that have polynomial length proofs (or solutions) which can be verified by a  $DLOG$  machine. In fact, the verification machine can also be a  $\mathcal{P}$  machine - this does not add to the power of the class. However,  $DLOG$  is preferable due to technical reasons. In terms of circuit,  $\mathcal{NP}$  corresponds to constant depth exponential size semi-unbounded circuit wherein the AND gates have polynomial fan-in. Note that the OR gates, by virtue of having unbounded fan-in can essentially have exponential fan-in. Important problems in  $\mathcal{NP}$  are Integer programming, Travelling salesman problem, Scheduling, VLSI routing and Satisfiability in propositional logic. All of these problems are  $\mathcal{NP}$ -complete.

(4) The class  $\mathcal{NCLOG}$  is the class that corresponds to a "one-way" proof given to a  $DLOG$  machine as auxiliary input (along with the regular input). It is the set of all problems that can be solved by polynomial sized circuit families that is composed only of OR gates. Important problems in this class are 2CNF satisfiability, Topological sort, source-destination connectivity.

(5) A very important parallel class is  $\mathcal{NC}$ . It is the set of all problems that can be solved by a family of polynomial sized circuits with poly-log depth. By poly-log we mean  $\log^{O(1)}(n)$ . Circuit depth is easily seen to model hardware time. If we assume that the circuit is nicely layered, we see that a parallel machine can evaluate the circuit layer by layer. This definition of complexity is very idealized. For example, communication costs and delays etc. are not taken into consideration. But if a problem cannot be solved in even such an idealized model, there is no hope of solving the problem on a parallel computer. Note that  $\mathcal{P}$  has circuits that could have polynomial depth. Important problems in  $\mathcal{NC}$  are sorting, merging, shortest paths, recognition problems concerning chordal graphs, orthogonalization of basis vectors, determinant, rank of matrix, eigenvalue computation. In fact,  $\mathcal{NCLOG}$  is a subset of  $\mathcal{NC}$ .

(6) In  $\mathcal{NP}$  instead of looking for a proof or solution to augment the input, we could also count the number of valid solutions. For example, the property of a graph having a Hamiltonian cycle is in  $\mathcal{NP}$ . The witness is a cycle in a graph which visits every vertex exactly once. Instead, we can count the number of hamiltonian cycles in the graph. In general the counting version of  $\mathcal{NP}$  problems form a class called  $\#\mathcal{P}$ . A very important problem in this class is that of computing the permanent. This problem is in fact complete for  $\#\mathcal{P}$ . It has already been mentioned that the determinant is in  $\mathcal{NC}$ , which is within  $\mathcal{P}$ . The class  $\#\mathcal{P}$  is now known to be a "big" class. So it is unlikely to have polynomial time algorithms. In this manner, complexity theory provides a new perspective to the determinant vs permanent question.

## 6. Conclusion

The main motivation for writing this paper is to create an awareness among scientists and engineers- that there is a very fundamental area of research in Computer Science, which is relevant to other branches of science and engineering.

The paper is informal in style and motivates the reader with many examples. The paper deals with the most elementary ideas in the field. Randomization is

mentioned but no randomized complexity class is introduced in the paper as the definitions are technical. Other advanced topics are outside the scope of the current paper. It should be added that the area has been around for twenty-five years and the last five years have seen some breathtaking results in this area.

## References

- Balcázar J L, Díaz J, Gabarró J 1990 *Structural complexity* (New York: Springer-Verlag) vols 1 & 2
- Cook S 1985 A taxonomy of problems with fast parallel algorithms *Inf. Comput.* 64: 2-22
- Garey M R, Johnson D S 1979 *Computers and intractability - A guide to the theory of NP-completeness* (San Francisco: Freeman)

# Managing interprocessor delays in distributed recursive algorithms

V S BORKAR and V V PHANSALKAR

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India

**Abstract.** For a class of distributed recursive algorithms, it is shown that a stochastic approximation-like tapering stepsize routine suppresses the effects of interprocessor delays.

**Keywords.** Distributed algorithms; tapering stepsize; stochastic approximation; strict Liapunov systems.

## 1. Introduction

In the distributed implementation of recursive algorithms, one often has to contend with the problem of interprocessor delays. There have been several attempts to model and analyse such phenomena mathematically, the earliest perhaps being the work of Chazan & Miranker (1969) on linear iterations. A comprehensive survey of the work up to late eighties, along with an extensive bibliography, can be found in (Bertsekas & Tsitsiklis 1989), chapters 6 and 7. In these references, one observes that the distributed algorithms are shown to work as desired only under rather strong conditions, such as boundedness of delays and other algorithm-specific restrictions. On a different note, the recent work of Gharavi & Anantharam (1992) analyses the complex behaviour that can result under stationary delays even in linear iterations.

The aforementioned works use a constant stepsize for deterministic algorithms. A tapering stepsize as in stochastic approximation theory by Benveniste *et al* (1990) is used only when there is a random 'noise' component and the aim there is to suppress the effect of noise on the asymptotic behavior of the algorithm. We argue here that the same scheme also suppresses the effect of delays. The main strength of our analysis is that we allow for very general delays, possibly unbounded, correlated and nonstationary, as long as a mild 'moment' condition is satisfied. Another important gain is that our analysis is not very algorithm-specific. It applies simultaneously to a large class of algorithms, viz., those whose continuous limit is a so called 'strict Liapunov system'. The latter subsumes a large number of gradient-like algorithms currently in vogue among the neural networks community (Hirsch 1987).

The paper is organized as follows. The next section describes the mathematical formulation of the problem and the assumptions. It also defines a strict Liapunov system and gives several examples of such systems. The third section proves our

main result. The last section concludes with some relevant remarks.

## Problem description

Consider a recursive algorithm in  $R^d$ ,  $d \geq 1$ , described by

$$X(n+1) = X(n) + a(n)F(X(n)) \quad (1)$$

here  $X(n) = [X_1(n), \dots, X_d(n)]' \in R^d$  with  $X(0)$  prescribed, and  $F(\cdot) = [F_1(\cdot), \dots, F_d(\cdot)]' : R^d \rightarrow R^d$  is a bounded map satisfying the Lipschitz condition:

$$\|F(x) - F(y)\| \leq K \|x - y\|; x, y \in R^d$$

for some  $K > 0$  (Prime denotes transposition of a vector or a matrix).  $\{a(n)\}$  is a prescribed sequence of strictly positive numbers. The  $i$ -th row of this vector iteration reads

$$X_i(n+1) = X_i(n) + a(n)F_i(X_1(n), \dots, X_d(n)). \quad (2)$$

In our model of distributed computation with delays, we replace (2) by

$$X_i(n+1) = X_i(n) + a(n)F_i(X_1(n - \tau_{i1}(n)), \dots, X_d(n - \tau_{id}(n))) \quad (3)$$

where  $\{\tau_{ij}(n), 1 \leq i, j \leq d\}$  are random delays encountered at the  $(n+1)$ -st iteration. Our aim is to analyse the system (3) under the following assumptions:

(A1)  $a(n) = 1/n, n = 1, 2, \dots, a(0) = 1$ .

(A2) The delays  $\{\tau_{ij}(n)\}$  satisfy: There exist  $b, C > 0$  such that

$$E[(\tau_{ij}(n))^b / X(m), m \leq n, \tau_{pq}(k), k < n, 1 \leq p, q \leq d] \leq C \quad (4)$$

with probability one, for all  $i, j, n$ .

(A3) The ordinary differential equation (ODE) described by

$$\dot{x}(t) = F(x(t)) \quad (5)$$

is a 'strict Liapunov system' described later in this section.

Some comments are called for regarding these assumptions. (A1) implies  $\sum_n a(n)^2 < \infty, \sum_n a(n) = \infty$ , which is the condition used in stochastic approximation theory to suppress the effect of noise. Our choice of  $\{a(n)\}$  coupled with (4) will simplify our analysis considerably. This choice, however, is not very rigid. One could use other stochastic approximation-like tapering stepsize routines in conjunction with an appropriate modification of (4), as will become evident later. Note that (4), in particular, implies that the conditional distributions of  $\tau_{ij}$  given the past are 'tight' in probability theory parlance. Finally, the Lipschitz condition on  $F$  implies that (5) has a unique solution for all  $t \geq 0$ . Later on we shall comment on the possibility of relaxing this condition.

A continuously differentiable function  $V : R^d \rightarrow R^+$  is said to be a Liapunov function for (5) if  $\nabla V \cdot F \leq 0$  everywhere and a strict Liapunov function if in addition,  $\nabla V \cdot F < 0$  outside the set  $E_F = \{z \mid F(z) = 0\}$ . Call (5) a strict Liapunov system if it has bounded trajectories,  $E_F$  consists of isolated points and a strict Liapunov

function  $V$  exists. (The latter would imply bounded trajectories if  $V(x) \rightarrow \infty$  as  $\|x\| \rightarrow \infty$ .) Some examples of such systems are:

(i) Gradient system: Consider  $F = -a\nabla f$  where  $a > 0$  and  $f: R^d \rightarrow R$  is a continuously differentiable, proper function (i.e.,  $f(x) \rightarrow \infty$  as  $\|x\| \rightarrow \infty$ ) with bounded gradient and isolated critical points.  $V(\cdot) = f(\cdot) + \min f(x)$  will do the job.

(ii) Newtonian systems (Vidyasagar 1978): For  $f$  as above, consider

$$\ddot{x}(t) + h(x(t))\dot{x}(t) + a\nabla f(x(t)) = 0 \quad (6)$$

where  $h: R^d \rightarrow R^+$  is bounded Lipschitz. This is Newton's law in a potential field with position-dependent kinetic friction. This can be rewritten as

$$\dot{x}(t) = y(t), \dot{y}(t) = -h(x(t))y(t) - a\nabla f(x(t)),$$

which conforms with (5). Assuming  $f \geq 0$ , the 'total energy'  $V(x, y) = 1/2y^2 + af(x)$  serves as a strict Liapunov function as long as  $h > 0$  outside  $E_f = \{x \mid \nabla f(x) = 0\}$ . For constant  $h$ , (6) will be recognized as the continuous analog of 'gradient descent with a momentum term' for minimizing  $f$ . For constant  $h$ , if we approximate  $f$  by a positive definite quadratic form near a local minimum of  $f$ , (6) looks like the familiar second order linear system of control systems (Ogata 1970) and may be expected to behave like one locally. That is, if  $h$  is chosen to make the latter 'underdamped', it will quickly hit a local minimum and then oscillate around it before settling down. The oscillations may be quenched by making  $h$  position dependent, with low values away from  $E_f$  and high values on  $E_f$ . Preliminary numerical experiments by the second author suggest that this is a promising strategy for speed-up.

(iii) Neural networks: The area of analog neural networks has several examples of strict Liapunov systems (Hirsch 1987; Schurmann 1989). Hirsch (1987) describes a scheme for building more complicated systems from 'cascades' of simpler ones, while Cohen (1992) considers the 'inverse problem' of constructing a strict Liapunov system with a prescribed equilibrium behaviour.

(iv) Fixed point algorithms: Recall the norms  $\|\cdot\|_p, p \in [1, \infty]$ , on  $R^d$  defined by: for  $x = [x_1, \dots, x_d]' \in R^d$ ,

$$\begin{aligned} \|x\|_p &= \left( \frac{1}{d} \sum_{i=1}^d |x_i|^p \right)^{1/p}, \quad 1 \leq p < \infty, \\ \|x\|_\infty &= \max_i |x_i|. \end{aligned}$$

Suppose  $F(x) = G(x) - x$  where  $G$  satisfies:

$$\|G(x) - G(y)\|_p \leq \alpha \|x - y\|_p, \quad x, y \in R^d,$$

for some  $\alpha \in (0, 1), p \in [1, \infty]$ . Then by the 'contraction mapping' theorem (Bertsekas & Tsitsiklis 1989),  $E_F$  is a singleton  $\{x^*\}$  (say). Then  $V(x) = \|x - x^*\|_p$  serves as a strict Liapunov function (Borkar & Soumyanath 1994). (One should note that  $V$  is not differentiable everywhere for  $p = 1, \infty$ , but this does not pose any problems, Borkar & Soumyanath 1994).

Our results will imply that discrete time, distributed implementations of (i) - (iv) above will asymptotically track the continuous dynamics under (A1) - (A2).

We conclude this section by recalling from Hirsch (1987) a result that will be central to our analysis. For given  $T, \delta > 0$ , a  $(T, \delta)$ -perturbed trajectory of (5) is a (possibly discontinuous) function  $y : [0, \infty) \rightarrow R^d$  such that the following conditions hold: There is an increasing sequence  $T_j \rightarrow \infty$  and solutions  $x^j(t), t \in [T_j, T_{j+1}]$  of (5) such that  $T_{j+1} - T_j \geq T$  for all  $j = 1, 2, \dots$ , and

$$\|y(t) - x^j(t)\| < \delta, \quad T_j \leq t \leq T_{j+1}, j \geq 1.$$

**Theorem 1.** *Hirsch (1987) for a strict Liapunov system (5), for any  $T, \epsilon > 0$ , there exists a  $\delta_0 = \delta_0(T, \epsilon) > 0$  such that if  $0 < \delta < \delta_0$ , then any limit point of a  $(T, \delta)$ -perturbed trajectory  $y(\cdot)$  is within  $\epsilon$  of a point  $p \in E_F$ . Moreover, if  $y(0)$  is in the domain of attraction of a stable equilibrium point  $q \in E_F$ , then for sufficiently small  $\delta$ , we can take  $p = q$ .*

The way we shall use this result is by showing that a suitably interpolated version of (3) is in fact a  $(T, \delta)$ -perturbed trajectory of (5). The details are carried out in the next section.

### 3. Main result

We shall proceed through a sequence of lemmas. Let  $T > 0$  and define time sequences  $\{t_n\}$  and  $\{T_n\}$  as follows:

$$\begin{aligned} t_0 &= T_0 = 0. \\ t_n &= \sum_{i=1}^n 1/i, \quad n \geq 1. \\ T_n &= \min\{t_m \mid t_m \geq T_{n-1} + T\}, n \geq 1. \end{aligned}$$

Then  $T_n = t_{m(n)}$  for a strictly increasing sequence  $\{m(n)\}$  and  $T_n \in [T_{n-1} + T, T_{n-1} + T + 1]$  for all  $n$ . Let  $I_n = [T_n, T_{n+1}]$ ,  $n \geq 1$ . Let  $x(t), t \in I_n$ , be a solution of (5) with  $x(T_n) = X(m(n))$  and define  $y(t), t \in I_n$ , by:

$$y(T_n) = x(T_n),$$

$$y(t_{m(n)+k+1}) = y(t_{m(n)+k}) + F(y(t_{m(n)+k}))(t_{m(n)+k+1} - t_{m(n)+k}),$$

with  $y(\cdot)$  linearly interpolated on the intervals  $[t_{m(n)+k}, t_{m(n)+k+1}]$ . Let  $\epsilon > 0$ .

*Lemma 1.* For sufficiently large  $n$ ,

$$\sup_{I_n} \|x(t) - y(t)\| < \epsilon.$$

*Proof.* For  $t \in I_n$ , let  $g(t) = \max\{t_{m(n)+k} \mid t_{m(n)+k} \leq t\}$ . Then for  $t \in I_n$ ,

$$y(t) = y(T_n) + \int_{T_n}^t F(y(g(s)))ds,$$



$$x(t) = x(T_n) + \int_{T_n}^t F(x(s)) ds.$$

Letting  $K, K'$  be resp. the Lipschitz constant of and a bound on  $F$ , we have

$$\begin{aligned} \|x(t) - y(t)\| &\leq K \int_{T_n}^t \|x(s) - y(s)\| ds + K \int_{T_n}^t \|y(s) - y(g(s))\| ds \\ &\leq K \int_{T_n}^t \|x(s) - y(s)\| ds + KK' \sum_{k=m(n)}^{m(n+1)} (1/k^2), \end{aligned}$$

where we use the fact that

$$\|y(t) - y(t_i)\| \leq K' |t_{i+1} - t_i| \text{ for } t \in [t_i, t_{i+1}].$$

The claim now follows from the Gronwall lemma (Vidyasagar 1978, pp. 292) and the summability of  $\sum (1/k^2)$ .  $\square$

Next, write (3) as a vector iteration

$$X(n+1) = X(n) + \frac{1}{n} Z(n)$$

with appropriately defined  $Z(n) = [Z_1(n), \dots, Z_d(n)]'$  and set

$$\hat{Z}(n) = E[Z(n)/X(m), m \leq n, \tau_{ij}(k), k < n, 1 \leq i, j \leq d], n \geq 0,$$

where the conditional expectation is componentwise. Let

$$M_n = \sum_{i=1}^n \frac{1}{i} (Z(i) - \hat{Z}(i)), n \geq 0.$$

*Lemma 2.*  $\{M_n\}$  converges as  $n \rightarrow \infty$ , with probability one.

*Proof.*  $\{M_n\}$  is seen to be a  $d$ -dimensional martingale (Neveu 1975) satisfying

$$\|M_{n+1} - M_n\| \leq 2K'd/n.$$

Thus,

$$E[\|M_{n+1} - M_n\|^2 / M_k, k \leq n] \leq K''/n^2$$

for a suitable  $K'' > 0$ . Since the right hand side is summable, the claim follows from Proposition VII-2-3(c), pp.149-150 (Neveu 1975).  $\square$

*Lemma 3.* There exist  $\bar{C}, r > 0, \bar{N} \geq 1$  such that for all  $n \geq \bar{N}$ ,

$$\|F(X(n)) - \hat{Z}(n)\| < \bar{C}/n^r.$$

*Proof.* Writing  $\hat{Z}(n) = [\hat{Z}_1(n), \dots, \hat{Z}_d(n)]'$ , we shall compare  $F_i(X(n)), \hat{Z}_i(n)$ , for a fixed  $i, 1 \leq i \leq d$ . Let  $1 > c > 0$ . We have

$$\begin{aligned} &|F_i(X(n)) - \hat{Z}_i(n)| \\ &\leq E[|F_i(X(n)) - Z_i(n)| I\{\tau_{ij}(n) \leq n^c \text{ for all } j\} / X(m), m \leq n, \tau_{pq}(k), \\ &k < n, 1 \leq p, q \leq d] + E[|F_i(X(n)) - Z_i(n)| I\{\tau_{ij}(n) > n^c \text{ for some } j\} / X(m), \\ &m \leq n, \tau_{pq}(k), k < n, 1 \leq p, q \leq d]. \end{aligned}$$

The second term can be bounded by  $2K'Cd/n^{bc}$  in view of (4) and the Chebyshev inequality. Let  $\bar{n}$  = the integer part of  $n^c$ . Let  $n$  be large enough so that  $n > \bar{n}$ . Then for  $m \leq \bar{n}$ ,

$$\|X(n) - X(n-m)\| \leq K' \sum_{k=n-\bar{n}}^n \frac{1}{k} \leq \bar{K}/n^{1-c}$$

for a suitable constant  $\bar{K} > 0$ . Thus the first term above may be bounded by  $K\bar{K}d/n^{1-c}$ . The claim follows.  $\square$

Now define  $\bar{X}(t)$ ,  $t \geq 0$ , by:  $\bar{X}(t_n) = X(n)$  for  $n \geq 0$ , with  $\bar{X}(\cdot)$  defined by linear interpolation on each internal  $[t_n, t_{n+1}]$ ,  $n \geq 0$ . For the next lemma, fix a sample point for which the conclusions of lemma 2 hold.

*Lemma 4.* For sufficiently large  $n$ ,

$$\sup_{I_n} \|\bar{X}(t) - y(t)\| < \epsilon.$$

*Proof.* Let  $n \geq 1$ . For  $i \geq m(n)$ , we have

$$\bar{X}(t_{i+1}) = \bar{X}(t_i) + \frac{1}{i}F(\bar{X}(t_i)) + \frac{1}{i}(\hat{Z}(i) - F(\bar{X}(t_i))) + \frac{1}{i}(Z(i) - \hat{Z}(i)).$$

Define  $\xi_i = M_i - M_{m(n)}$ ,  $\hat{X}_{i+1} = \bar{X}(t_{i+1}) - \xi_i$  for  $i \geq m(n)$  with  $\hat{X}_{m(n)} = X(m(n))$ . (Thus  $\xi_{m(n)-1} = 0$ ). Then

$$\hat{X}_{i+1} = \hat{X}_i + \frac{1}{i}F(\hat{X}_i) + \frac{1}{i}(F(\hat{X}_i + \xi_{i-1}) - F(\hat{X}_i)) + \frac{1}{i}(\hat{Z}(i) - F(\bar{X}(t_i))).$$

Also,

$$y(t_{i+1}) = y(t_i) + \frac{1}{i}F(y(t_i)).$$

Thus for  $n$  sufficiently large, lemma 3 leads to

$$\|\hat{X}_{i+1} - y(t_{i+1})\| \leq (1 + \frac{K}{i}) \|\hat{X}_i - y(t_i)\| + \frac{K}{i} |\xi_{i-1}| + \frac{\bar{C}}{i^{1+r}}.$$

Let  $\delta > 0$ . By lemma 2, for sufficiently large  $n$ ,

$$\sup_{i \geq m(n)} |\xi_i| < \delta/2,$$

$$\sum_{i \geq n} (1/i^{1+r}) < \frac{\delta}{2}.$$

Thus for  $n$  sufficiently large, using the inequality  $1 + \frac{K}{n} \leq e^{K/n}$  and iterating, one has

$$\sup_{m(n) \leq i \leq m(n+1)} \|\hat{X}_i - y(t_i)\| \leq e^{K(T+1)}(\bar{C} + K(T+1))\delta.$$

Since

$$\sup_{m(n) \leq i \leq m(n+1)} \|\hat{X}_i - \bar{X}(t_i)\| < \delta/2,$$

the claim follows on choosing  $\delta$  sufficiently small and observing that  $\bar{X}(\cdot), y(\cdot)$  are linearly interpolated from their values at  $\{t_i\}$ .  $\square$

We are now ready to prove our main result.

**Theorem 2.** *The iteration (3) converges with probability one to a single point in  $E_F$ . Furthermore, the iterates visit the basin of attraction of a stable equilibrium point in  $E_F$  infinitely often if and only if they converge to the same.*

*Proof.* From lemmas 1, 4 and theorem 1, it follows that the limit points of the iteration (3) are in  $E_F$ . The fact that the stepsize  $a(n)$  is monotonically decreasing to zero ensures that the set of limit points is connected. Since  $E_F$  consists of isolated points, the set of limit points must therefore be a singleton. The first claim follows. The second claim follows from the second half of theorem 1.  $\square$

Note that the property of a point being in the basin of attraction of a stable equilibrium point is generic for strict Liapunov systems. Thus the behaviour described in the second half of the theorem is generic.

#### 4. Concluding remarks

We conclude by pointing out additional advantages of our model and some possible generalizations.

(i) Note that our model does not require that each value of  $i - th$  component computed at some stage be transmitted to the processor computing  $j - th$  component, or that the latter use it even when he receives it. The only requirements are that (a) if computation of  $i - th$  component requires the previous value of  $j - th$  component, then the latter be transmitted to  $i$  infinitely often, (b) the received value of the same, if used by  $i$ , be used without too much delay, i.e., such that (A2) holds. Of course, this concerns only the convergence of (3), not the speed thereof, which is a far more delicate issue that needs further study.

(ii) Several of our assumptions could be relaxed at the expense of additional technicalities. We have already mentioned that other choices of  $\{a(n)\}$  with  $\sum_n a(n) = \infty, \sum_n a(n)^2 < \infty$ , could be explored. Boundedness and Lipschitz conditions on  $F$  could be dropped to some extent. For example,  $F$  being locally Lipschitz with at most linear growth will suffice for the global existence of a solution to (5). Alternatively, the condition that  $V$  is proper will do. The use of Lipschitz condition in our proofs can be replaced by locally Lipschitz property if we show a priori the fact that the iterates of (3) remain in a bounded set with probability one. Stochastic Liapunov arguments may help here. Finally, the proof of theorem 1 in Hirsch (1987) seems to extend easily to the case when  $E_F$  is not necessarily a collection of isolated points, but contains connected sets of local minima any two of which are separated by an amount exceeding a fixed  $\Delta > 0$ . Hirsch (1987) does not deem this situation interesting because  $E_F$  being a set of isolated points is generic from structural stability point of view. Nevertheless, one does encounter the more general situation mentioned above in 'learning' and 'identification' applications because of overparametrized model sets and therefore merits attention.

(iii) In stochastic approximation (Benveniste *et al* 1990), one studies iterations of

the type

$$X(n+1) = X(n) + a(n)G(X(n), \xi(n)) \quad (7)$$

with a prescribed  $X(0)$ ,  $\{a(n)\}$  as above, and  $\{\xi(n)\}$  (in the simplest case) is a sequence of independent and identically distributed random variables, interpreted as 'noise'. Under mild conditions, one shows (Benveniste *et al* 1990) that this iteration tracks the continuous dynamics (5) with  $F$  defined by

$$F(x) = \int G(x, y)\mu(dy),$$

$\mu$  being the law of  $\xi_1$ . If this  $F$  satisfies the conditions of this paper, our analysis extends to the distributed version of (7) with delays as in (3), satisfying (A2). This requires very minor modifications in the proof. An important special case is when  $G(x, y) = -\nabla f(x - y)$  with  $f$  as in example (i) of section 2. If  $\mu$  has a density  $\varphi$ ,  $F$  becomes

$$F(x) = - \int \varphi(x - y)\nabla f(y)dy. \quad (8)$$

If  $\varphi$  is a 'nice' i.e. sufficiently smooth function,  $F$  may be viewed as a low pass filtered version of  $\nabla f$ , or equivalently, gradient of a low pass-filtered version of  $F$ . If  $F$  is jagged with several small, sharp valleys, it is not amenable for gradient-based methods and the right dose of low pass filtering may be expected to replace it by a better behaved function without significantly perturbing the location of Argmin ( $f$ ). In this viewpoint,  $\{\xi_n\}$  is not 'noise', but a benign randomization that may be deliberately introduced to achieve the same effect as the computationally expensive convolution in (8). One may hope to further improve on the performance of this algorithm by progressively decreasing the extent of low pass filtering, e.g., by replacing the above algorithm by

$$X(n+1) = X(n) + a(n)(-\nabla f(X(n) - b(n)\xi(n)))$$

where  $\{\xi(n)\}$  are i.i.d., zero mean and  $\{b(n)\}$  decrease to zero at a rate slower than that of  $\{a(n)\}$ . This is closely related to a special case of a broad class of algorithms studied by Gelfand & Mitter (1992). These algorithms asymptotically track not an ODE as in (5), but a time-inhomogeneous stochastic differential equation. (In fact, the latter is precisely the Langevin algorithm of continuous time simulated annealing.) It would be interesting to extend our analysis to algorithms of this type.

The research of the first author was supported by the grant ISTC/EE/VS/049 from the ISRO-IISc Space Technology Cell. The research of the second author was supported by U.S. Office of Naval Research Grant No.00014-92-J-1324 under an Indo-U.S. project.

## References

- Benveniste A, Metivier M, Priouret P 1990 *Adaptive algorithms and stochastic approximations* (Berlin-Heidelberg: Springer-Verlag)
- Bertsekas D P, Tsitsiklis J N 1989 *Parallel and distributed computation* (Englewood Cliffs, NJ: Prentice Hall)
- Borkar V S, Soumryanath K 1994 A new analog parallel scheme for fixed point computation, part I: theory. Preprint
- Chazan D, Miranker W 1969 *Chaotic oscillations, linear algebra and its applications*. 2: 199-222
- Cohen M A 1992 The construction of arbitrary stable dynamics in neural networks. *Neural Networks* 5: 83-103
- Gelfand S B, Mitter S K 1992 Recursive stochastic algorithms for global optimization in  $R^d$ . *SIAM J. Contr. Optim.* 29: 999-1018
- Gharavi R, Anantharam V 1992 A structure theorem for partially asynchronous relaxations with random delays. Preprint
- Hirsch M W 1987 Convergent activation dynamics in continuous time networks. *Neural Networks* 2: 331-349
- Neveu J 1975 *Discrete parameter martingales* (Amsterdam: North-Holland)
- Ogata K 1970 *Modern control engineering* (Englewood Cliffs, NJ: Prentice Hall)
- Schurmann B 1989 Stability and adaptation in artificial neural systems. *Phys. Rev.* A40: 2681-2688
- Vidyasagar M 1978 *Nonlinear systems analysis* (Englewood Cliffs, NJ: Prentice Hall)



# Traffic engineering in ATM networks: Current trends and future issues

A Gravey<sup>1</sup>, G Hébuterne<sup>2</sup>, R R Mazumdar<sup>3</sup> and C Rosenberg<sup>4</sup>

<sup>1</sup> France Telecom, CNET-Lannion A, Route de Trégastel, 22300 Lannion, France

<sup>2</sup> Département RST, Institut National des Télécommunications, 91011 Evry Cedex, France

<sup>3</sup> INRS-Télécommunications, Université du Québec, Ile-des-Soeurs, P.Q. H3E 1H6, Canada

<sup>4</sup> Département de Génie Electrique, Ecole Polytechnique, Montréal, P.Q. H3C 3A7, Canada

**Abstract.** In this survey we present an overview of the current state-of-the-art and the future issues to be resolved for the deployment of integrated communications networks based on the Asynchronous Transfer Mode (ATM). Our presentation is from the perspective of traffic and congestion control and management issues which hold the key to the successful achievement of this deployment since it is mainly in these aspects that ATM networks differ from conventional communication networks.

**Keywords.** ATM; Traffic characterization; control; quality of service.

## 1. Introduction

In the past the evolution of telephone and data networks has gone ahead independently, the former was performed by the telephone companies while the latter was done by the computing community (the ARPAnet being the case in point). This has basically involved the setting up of a parallel infrastructure or communications backbone suited to the needs of the traffic characteristics of each service. With the need for economic efficiency and a convergence in the demands from both directions to provide a reliable, high bandwidth communications backbone to offer enhanced or new services, there was a consensus in the seventies to develop Integrated Services Digital Networks (ISDN) which would provide a common access for data as well as voice traffic. The ISDN concept was based upon a digital telephone network using common channel signalling to allow basic access rates based upon multiples of 64 kbits/s which is the fundamental building block of ISDN services. However real integration was not performed since the two traditional transfer modes were coexisting just behind the access point.

With the explosion in the communication needs in the eighties driven by the need to provide access to traffic requiring much higher bandwidth (bitrates) than was anticipated the transition towards Broadband ISDN (B-ISDN) networks was made. The idea of B-ISDN is very revolutionary compared to ISDN since the goal is to provide not only a common access to all existing and future communication services but also a common transfer mode to allow an efficient and transparent transport of information while assuring a service dependent Quality of Service (QoS) which is usually specified in terms of delay and/or loss constraints.

Due to the extremely high speeds required for some applications such as real-time video and medical imaging it was quickly realized that the conventional 7-layer ISO architecture on which ISDN is based as well as the error and congestion control procedures existing in layer 2 and 3 would be the limiting bottleneck. This is due to the fact that many of the control procedures associated with traffic management procedures in a X.25-like network such as link-by-link error control would inhibit the speeds and are not necessary thanks to the use of virtually error-free optical fiber transmission links. Besides the ability to react to congestion through feedback would be impossible since by the time the information is fed back the state of the network could be vastly different making the information completely useless. Thus, it was natural for introducing a change in paradigms in the way traffic management and control is to be done.

The shift in paradigms meant that to achieve high speeds it is essential to reduce the overheads by the migration of "intelligence" towards the perimeter or input nodes while providing a "transparent" high-speed and reliable communication link between user and destination. In addition the network architecture should incorporate the flexibility of providing the nice features of circuit-switched networks and the flexibility and efficiency of packet switched networks. These characteristics are to be found in the concept of the Asynchronous Transfer Mode (ATM), originally proposed by CNET in France, which was adopted as the standard by the ITU-T (successor to the CCITT) in 1989 (I.121 1989). The basic concept of ATM is to provide a powerful inter-connection medium with limited functionalities ("intelligence") but offering high performance in terms of bandwidth, QoS and access. This would allow for the means to integrate all existing services and offer high bit-rates to end users through a paring down of the number of layers of the protocols.

In order to appreciate the issues it is worth examining the current services that are envisaged and why the traffic engineering issues are central to the design of ATM networks. Before we do so we state a few provisos: this paper is not an introduction to ATM networks (for an introduction to ATM see CNET (1991) and Minzer (1989)) rather this paper highlights the traffic engineering issues with a presentation of the basic problems, current trends and future issues to be resolved if the goal of providing a fully integrated, versatile and reliable communications network is to be realized. The second point is that we do not survey all the trends but restrict ourselves to the issues in the context of standards that have been agreed upon since if future research is to have any impact it must be within the guidelines of the standards. Finally the issues presented here are not exhaustive and research in this area is being pursued actively the world over.

Let us begin with standard telephone traffic. The analog voice signal is sampled and digitized, resulting in a continuous bit stream: 1 byte every  $125\mu\text{s}$ . The main constraints to handle such a service are: no variation in the transport delay of



successive bytes (or at least a very limited variation in case of packetized voice), low delay (say around 20ms to avoid the need for echo cancelling), allowable losses upto ratios around  $10^{-3}$ , low cost (in terms of complexity) of end equipment.

Data transfer is another kind of service, with different characteristics: information arrives in long blocks (from kbytes to Gbytes in the case of file transfers); there is usually no constraint on the delay or on its variation; loss must be kept minimum – less than  $10^{-9}$ ; the complexity of existing end equipment eases implementing elaborate procedures.

Naturally, all intermediate cases could be imagined. Video for professional or entertainment usage; high quality sound; etc. In each case, the service may be characterized in terms of delay or loss constraints and of bandwidth requirements. It is characterized also by the statistical description of the data flow to be carried. Note that most services to come will be *multimedia*, meaning a mix of basic services in a single offer.

The coexistence of these services puts strong requirements on the network. This has been summarized under the concept of *transparency*: the source generates its traffic at its own rate, and ideally the network should carry the traffic at the same rate and without any alteration.

In order to fulfill the above requirements, all services in the B-ISDN are offered above a common medium, the ATM network, that provides basic and powerful interconnection facilities. Information is segmented into fixed-size *cells* at network input nodes. The cell consists of a 48-byte information field and a 5-byte header. Cells are carried along *Virtual Path Connections* (VPC) or *Virtual Circuit Connections* (VCC), identified by the VPI- and VCI-field of the header. The ATM layer performs only routing operations. No error detection/correction takes place in the network (note however that the header is protected by a 1-byte CRC field: cells with corrupted headers are discarded).

The ATM layer of the B-ISDN performs the basic transport function in a simple and efficient way. All its complex procedures are hardware-implemented, yielding a high performance level and the capability to operate at high bandwidths. Specific additional functionalities, such as jitter or error correction, are to be offered on a per-service basis, by an ad-hoc “*ATM Application Layer*” (AAL). Thus we can view the ATM layer as a fixed and short size cell relay network.

Note that the existence of VPC as well as VCC allows for a wide range of semi-permanent and switched accesses.

This organization aims at giving the B-ISDN the higher “flexibility” – ease in introducing new services with specific Quality of Service constraints.

However, a problem arises concerning *traffic and congestion control*. With traditional circuit switched networks, a source cannot exceed the basic rate offered to it (64 kbit/s, in telephone network). Most of the time, the ATM network will offer the source full access to a 150 Mbit/s link. In X.25-like packet networks, link-by-link congestion control and acknowledgments prevent a source from flooding the network with its traffic. The lack of elaborate protocols in the ATM layer prevents the network management from relying on such procedures. In addition the sheer speeds makes the introduction of state dependent control mechanisms of limited value except in the case of congestion notification etc. and at the stage of admitting a connection. Hence, traffic control procedures must rely on “open-loop” measures, i.e., ensuring that traffic inputted into the network behaves “well”. This

basically implies that the traffic control mechanisms have to do the job of verifying the conformance of each traffic stream with the amount of resources allocated to the connection and the state of the network.

The aim of this paper is to point out how traffic and congestion control happens to be the central issue in ATM networks and the current approaches to address these problems.

The organization of this paper is as follows: In §1 we describe the challenge of traffic and congestion control and in particular, we discuss the issue of source characterization and the Quality of Service, source policing, Call Acceptance Control which is the procedure by which a call is accepted and the appropriate bandwidth allocated. In §2 we discuss the present status of traffic control in ATM networks. In Section 3 we discuss the issues which still require further studies like the one related to the choice of a multiplexing scheme to offer appropriate bandwidth which is aimed at increasing network efficiency.

## 2. Challenge of traffic and congestion control

One of the basic features of ATM networks is that at call set-up a negotiation takes place between the source and network. The source is supposed to indicate its characteristics and requirements in terms of QoS (loss probabilities, delays etc.) and the network decides whether to accept the connection. Based on the QoS and source characteristics the network makes decisions to allocate the appropriate bandwidth and buffers etc. and the network is committed to deliver the QoS if the the sources conforms with respect to its characterization. Of course due to the stochastic variation in the cells emitted by a source there could be a potential gain to statistically multiplex the different sources to gain efficiency of goodput (throughput meeting QoS). For this an efficient characterization of the source is needed.

For a source to declare its complete statistical characterization is completely unreasonable and besides the statistics could be altered (especially in the case of VPC) so one has to seek robust techniques for source characterization with a minimal amount of information to be conveyed. In addition such a minimal characterization can only hold in some approximate sense (i.e., with a closing error) since it is unreasonable to presume that a source will obey the characteristics conveyed in an absolute sense.

Once resources are committed on acceptance of a connection, the network must ensure that the source conforms to the declared characteristics. This control function has been termed source policing and lack of such a control will affect the QoS of background connections when the network tries to maximize efficiency.

According to the ITU-T (I.371 1993), a source submitting a connection request must provide a set of parameters to the network, the Source Traffic Descriptor, whose contents are meant to capture the source intrinsic behaviour. These parameters are part of a traffic contract negotiated between the user and the network. The *Source Traffic Descriptor* (STD) is a subset of the ATM Traffic Descriptor; in addition to traffic parameters, the contract also covers Quality of Service objectives and Cell Delay Variation tolerances. The declared traffic parameters must then be enforced (policed) by the network to protect the grade of service offered to compliant connections.

Hence one of the great challenges is to give an adequate characterization of source

traffic based on a minimal description. The source characterization must also offer maximum flexibility to the source in the way its traffic is generated.

## 2.1 Characterizing source traffic

There are various ways to describe the cell stream issued by a source. On the one hand, one can classify the sources according to their statistical pattern; on the other hand, "significant" parameters must be defined, to quantify the corresponding traffic process.

Significant parameters should measure the way the given source may influence the network behaviour (i.e., the impact on other already set-up connections), and should allow us to compute the QoS the network may offer the new connection. A typical scenario for the selection of source traffic parameters is in the framework of figure 1. All sources share a common communication resource, and are multiplexed through a finite-size buffer. The STDs must allow us to estimate the resulting QoS, typically loss probability and delay.

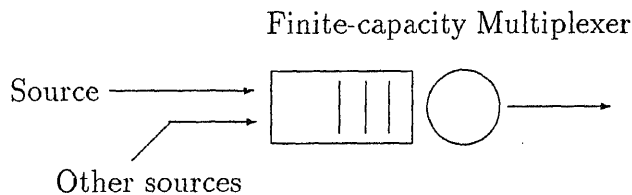


Figure 1. The queueing model of multiplexing sources

An important parameter arises naturally, it is the *peak cell rate*. This is the maximum rate at which the source is able to send data and can be viewed as the minimum interval between two consecutive cells. Its formal definition in the ITU-T framework calls for the *Equivalent Terminal* formalism (I.371 1993). Standardization efforts at the ITU-T have focused so far only on issues related to the peak cell rate.

Before we discuss the problem of source characterization beyond the peak cell rate, it is important to note that as a first cut the following categories of sources are naturally identified.

*Constant bit rate source (CBR):* This is the simplest case; the source generates a cell periodically, at its cell peak rate. The usual telephone traffic belongs to this class (64 kbit/s is 1 cell every 6 ms).

*Variable bit rate source (VBR):* The simplest video transmission would generate CBR streams. However, video coders efficiently compress the information stream, resulting in variable bit rates. The VBR source is described as having a long-term bandwidth requirement significantly lower than its peak cell rate, with an instantaneous demand varying continuously. Many applications can be viewed as VBR sources (interactive data processing, for instance).

*Sporadic source (on/off source):* This type of VBR source corresponds to a two-state CBR source. During the "On" state, the source generates a continuous stream.

During the "Off" state, the source remains silent. This definition can be generalized by allowing the process to switch between several CBR levels. Usually, the sojourn in each state has a random duration. One can give numerous examples of this kind of source. File transfer and still picture transmission (medical imaging) are the simplest representatives. For such applications, the peak cell rate of the "On" state is generally quite higher than the mean (it conditions the duration to transmit the picture).

Note that a "good" traffic parameter should be intrinsic, enforceable and usable in the allocation process. Various parameters have been proposed to the purpose of source characterization beyond the peak cell rate to allow statistical multiplexing. The first one is the *mean bit rate*. Roughly, this is the rate at which cells arrive, estimated over a long period of time. Note that only a Poisson arrival process is fully characterized by this parameter. Note also that enforcing such a parameter would, in general, require measures on a very long period.

The *burstiness* is the ratio of the peak to the mean cell rates. It is introduced as a way to quantify the variability of the arrival process.

Other ways can be used to describe the variability, and the correlation of the cell flow. For instance, let  $N(t)$  be the number of cells arriving in  $(0, t]$ . One can compute the mean and variance of  $N(t)$ , and the "*Index of Dispersion for Counts*" is defined as  $I(t) = \text{Var}[N(t)]/E[N(t)]$ . One should choose one point of  $I(t)$ , for instance the maximum value, or the limit as  $t \rightarrow \infty$ , etc.

These parameters undoubtedly present interesting features; unfortunately, it is not clear how they can be used to allocate resources in a statistical multiplexing context and how they could be efficiently enforced.

Another approach of the source characterization consists in postulating a source process of the type "*n-stage Markov Modulated Bernoulli Process*" (MMBP<sub>n</sub>): in each stage, cells are sent according to a pure chance process (Bernoulli), and jumps between states occur according to a Markov process. The parameters are chosen such that the MMBP "looks like" the actual process.

It must be remarked that, for a statistically well-defined class of sources, efficient traffic descriptors can be found. For instance, an On/Off source is perfectly characterized by the length of the emission and silence intervals; the Poisson arrival process is fully characterized by the parameter " $\lambda$ ". Unfortunately, this kind of characterization is of little value since the network would have to identify both the class and the corresponding parameters which could be only done theoretically with an infinite record.

To circumvent these difficulties an approach that has been adopted within the standardization bodies is to search for algorithmic rules i.e., rules such that if the cell count confirms according to some parameters within a prescribed time-interval then the source is said to be conforming. A second and more recent approach is via the use of the notion of "effective bandwidths" which tries to identify a single parameter associated with a source such that if bandwidth greater than the effective bandwidth is allocated the cell stream will meet the required QoS. We briefly discuss these issues and a more recent approach due to the authors based on quantile measures.

*Algorithmic descriptors:* Recently an algorithmic procedure called the Generic Cell Rate (GCRA) orum. This is very close to the  $(\sigma, \rho)$  characterization introduced by

Cruz (1991). These methods are in fact a particular instance of a broader class of algorithmic characterizations as shown by Berger (1994). The idea is that an arrival stream is said to be  $(\sigma, \rho)$  constrained if in the interval  $[t_0, t]$

$$N[t_0, t] \leq \rho(t - t_0) + \sigma$$

Suppose that the traffic stream is  $(\sigma, \rho)$  constrained for all  $t$ , then it is easy to see that from the above form the mean rate is bounded by  $\rho$  and  $\sigma$  is a measure of tolerance (a measure of variability) of the stream. Of course it is difficult a priori for a source to give such a characterization but such a characterization is directly enforceable by a simple *leaky bucket* (see discussion below in the context of control). These descriptors have gained much acceptance due to their simple characterization and the means to enforce the parameters. However such a characterization does not yield tight performance bounds since it is essentially a deterministic measure which gives "worst-case" bounds. Much work needs to be done on obtaining tight bounds from which we can study the multiplexing gains when several streams are multiplexed on a common line.

*Effective bandwidths:* The notion of an effective bandwidth of a source was first introduced by Hui (1988). The idea is to specify a single parameter called the effective bandwidth such that if the effective bandwidth is less than the multiplex rate (bandwidth allocated) the QoS constraints are met. This idea has been further refined by Kelly (1991) for shared resource systems and in the context of buffered resources by Elwalid & Mitra (1993). More recently a detailed exposition of this idea and its relation to the so-called rate functions in large deviations theory has been presented by Whitt (1995). By now the idea is fairly well understood for buffered and unbuffered systems when viewed as systems in isolation with a given mix of traffic. We briefly discuss this idea below.

Suppose there are  $N$  types of ON-OFF sources whose ON distributions are specified by  $F_i(x)$ ,  $i = 1, \dots, N$ . Suppose the sources wish to access a common multiplex of bandwidth  $C$  and a given source requires one unit of bandwidth when ON. Suppose we wish to allocate bandwidth such that the loss probability is less than  $e^{-\gamma}$ . Then it has been shown by Hui that there exists a parameter  $\alpha_i(F_i)$  which depends on  $C$  and  $\gamma$  such that if

$$\sum_{i=1}^N \alpha_i(F_i) < C$$

then the sources will suffer blocking less than  $e^{-\gamma}$ . Kelly (1991) showed this also holds for delay constraints. The parameter  $\alpha_i$  is called the *effective bandwidth* and represents the bandwidth requirements of the source to meet its QoS constraint. Note  $\alpha_i(F_i) \leq 1$ .

One of the important issues to resolve is given that there is a new request for a connection can one compute the effective bandwidth of the new connection without recalculating the effective bandwidths of all existing connections and the effects of the scheduling policies? Recent work on the notion of "decoupling bandwidths" by Di Vecinia *et al* (1994) is a step towards this direction but appears overly conservative. Moreover in the buffered resources context, the effective bandwidth is asymptotic as the buffer size increases and thus reduced utility since large buffers

will contribute to longer delays. Another issue concerns the enforcement of such a parameter. Nevertheless this does not mean that such a concept is not useful.

*Time-quantile measures:* While characterizing source behavior with one or a few parameters is a worthy goal it is of interest to ask the question what statistic associated with the arrival counting process (other than the obvious true interarrival distributions) characterizes the process from the point of view of performance and control objectives? In recent work Rosenberg *et al* (1993) showed that the so-called time- $\epsilon$ -quantile process yields such a useful criterion. The time-quantile process associated with a point process is the function, denoted  $M_\epsilon(T)$  which corresponds to the  $(1 - \epsilon)$  quantile for the number of counts in an interval of length  $T$ . If  $\epsilon$  corresponds to the loss level required such a function determines the maximum number of arrivals in an interval of length  $T$  with prob  $(1 - \epsilon)$ . It can be shown that statistics such as the indices of dispersion, mean and burst rate can be extracted from such a measure which can be determined empirically if the source is stationary in the long term [?]. Moreover this measure is directly enforceable. If one defines the function  $r(T)$  corresponding to the maximum number of cells accepted by a leaky bucket in any interval of length  $T$  then a cell loss rate  $< \epsilon$  is assured if  $M_\epsilon(T) < r(T)$  for every  $T$  (Rosenberg & Lague 1994)

In spite of the above developments traffic characterization remains an important issue since the gains of statistical multiplexing can only be achieved by a tight characterization which may imply complex procedures at the user end for determining it and heavy signalling to pass the information to the network. Moreover, in some remarkable empirical measurements on the aggregated LAN traffic reported by Leland *et al* (1994) it has been suggested that such traffic has fractal characteristics. This renders much of the issues such as effective bandwidths moot since the long range dependence makes the distributions heavy tailed and hence large deviations approaches might fail. Recently, Norros (1994) studied a model based on a simplification that the fractal traffic is modeled as a shifted fractional Brownian motion. Even based on this approximation he showed that the loss characteristics are crucially dependent on the *Hurst parameter* (which governs the rate of decay of the covariance and is a measure of the fractal dimension) and furthermore that having long buffers does not improve the situation and thus this type of traffic must be included into the categories of traffic models.

For a comprehensive survey of source characterization, the reader is referred to Roberts (1991) and the recent proceedings of the International Teletraffic Congress (Labetoulle & Roberts 1994).

## 2.2 Open-loop control of a source

As already mentioned, in the ATM context feedback or reactive control of sources is not feasible due to the high delay-bandwidth product. This is due to the fact that the delays in feeding back the status of the links are of comparable order to the transmission times so that by the time status information is received the state could be drastically different making the information useless. Hence the way the control function is envisaged is via open-loop or pro-active means by which desirable behaviour is enforced at the input point. In the ATM context this is referred to as

source policing. The requirements of such a control mechanism are that the function must be easy to implement and must react fast to non conforming behavior while being transparent to conforming streams.

The problem could be mathematically posed as follows: Let  $\varphi(\cdot)$  denote the policing control which acts on the input stream  $A(0, t)$  yielding an output stream  $D(0, t)$  i.e.,

$$\varphi(A(0, t)) = D(0, t)$$

Then the requirement is that  $D(0, t)$  which is the policed input to the network must be such that it conforms with the negotiated contract to meet the QoS and  $\frac{D(0, t)}{t}$  should be maximal i.e., the amount of throughput should be maximized.

It turns out that the map  $\varphi(\cdot)$  which is optimal is the so-called *leaky bucket* (LB for short) mechanism which we describe below: consider a *virtual queue* in which service tokens are generated at constant rate  $R$  and an arriving cell consumes a service token. There are a maximum of  $N$  tokens available and thus if on arrival there are no tokens available then the corresponding cell is either dropped or "tagged" depending on the philosophy of the network (see below). Note when service tokens are available the cells transit through the LB without incurring any delay.

Let  $r(T)$  denote the maximum number of cells accepted in an interval of length  $T$  by the mechanism ( $r(T)$  is a deterministic upper-bound on the number of cells allowed by the LB to enter the network over any time interval of length  $T$ ). Then it is easy to see that  $r(T) \leq N + RT$  or it is  $(N, R)$  constrained at the output. The interpretation is that  $N$  is the maximum instantaneous burst allowed and  $R$  the bound on the long-term average.

This is a particularly simple mechanism and was proposed by Turner (1986) in his seminal paper. The optimality of the mechanism in the sense that it is the mechanism which maximizes the throughput subject to the  $(N, R)$  constraint on the output stream was recently shown by Anantharam & Konstantopoulos (1994).

ITU-T has defined an algorithm which generalizes the LB: it is the *Virtual Scheduling Algorithm* (VSA) – or Continuous State Leaky Bucket (I.371 1993). It can be seen as a LB where the parameters  $N$  and  $T = 1/R$  would not be restricted to integer numbers.

In spite of the simple form of the control mechanism the problem of controlling a source is far from trivial which we discuss below.

Ideally, a Source Traffic Descriptor (STD) should be controlled at the interface point between the source and the network. Actually, the control process may not be performed at such a point, most of the time: as illustrated on figure 2, either the source is connected by means of a customer equipment (e.g., a local network), or the control is performed at an Inter-Carrier Interface (ICI). As a consequence, the control device only measures the source process "corrupted" by one or more exogeneous processes. Some parameters have to be set in the traffic contract, in order to take into account the modification of the STDs in the customer equipment.

Of primary concern here is the perturbation on the cell arrival process of the source to be controlled. Let us assume, for simplicity, that a CBR source is observed. Cells ought to be detected at epochs of the form  $T_0 + kT$  ( $T$  is the emission period). Most of the time, cells encounter some *Cell Delay Variation* (also called *jitter*) – that is they arrive sooner or later as compared with theoretical epochs. Possibly, one may even observe *bursts* of cells (i.e., groups of back to back cells). Since the phenomenon is inherent to an asynchronous transfer mode, this is the task of the control process

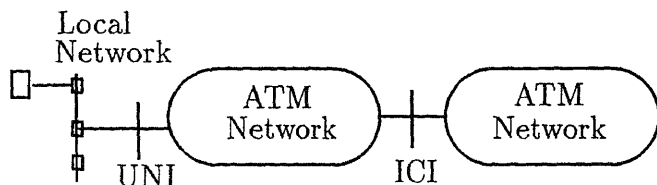


Figure 2. The general configuration: the Source is not connected directly at the UNI

to take it into account. The influence of the jitter must be upperbounded and the tolerance of the control algorithm must be tuned accordingly.

It has been recognized that cell jitter is composed of two basic complementary phenomena, namely cell clumping and cell dispersion (Roberts & Guillemin 1992). To illustrate the cell clumping effect, consider an initially periodic cell stream with period  $T$  passing through an ATM network. When this cell stream is observed at some point along the connection it may happen that the distance between two consecutive cells is less than the initial period  $T$  because of random delays. This is precisely what is called cell clumping. Its impact on the ATM layer performance is obvious since by creating bursts, it may affect the Cell Loss Ratios and thus must be controlled for the ATM layer sake (this has been realized a long time ago and cell clumping is taken care of in I.371 (see below)). It may also happen that two consecutive cells experience increasing delays so that their inter-arrival time at the observation point is greater than the initial period  $T$ , giving rise to a cell gap in the cell stream. This is the phenomenon of cell dispersion, which is critical for connections supporting a circuit emulation (e.g., ATM Adaptation Layer, AAL, of type 1) and/or when resequencing (e.g., two-layer video) has to be performed in the receiving terminal.

Corresponding to each control device, one can define the *worst-case* traffic. This is the "worst" source that is observed as conforming to the negotiated contract. By "worst source" is meant here the source inducing the higher loss ratio for its companion flows in any queue, like the one of figure 1. The safest approach is then to allocate resources based on the worst case traffic, the general belief concerning the worst case traffic for a LB is that the worst case is given by an On/Off periodic source filling the bucket with a burst at its peak rate and then remaining silent during a time just long enough to regenerate all the credits. For instance, assume one intends to control the peak rate of a CBR source, with a Leaky-Bucket, of parameters  $(R, N)$ . The worst-case source corresponding to the  $(R, N)$ -LB is an On/Off process, with On periods generating  $N/(1 - R)$  cells and silent periods of duration  $N/R$ . The network accepting the CBR source must be ready to carry in fact such an On/Off process. See Doshi (1993) for more considerations about the meaning to be given to the concept of "worst case source".

Numerous studies have been done (see Rathgeb 1991) concerning the capabilities and relative performances of these mechanisms. Is it possible for instance to control the mean rate – say  $m$  – of a connection? Let us try a LB with  $R$  slightly higher than  $m$ . The value of  $N$  has to be chosen such as to compensate for statistical fluctuations in the cell generation process which implies in general the use of a large  $N$  which results in poor response time and protection against bursts. Similarly, controlling the peak rate  $D$  would require  $R$  around  $D$ , and theoretically  $N = 1$



would be sufficient if there was no jitter but  $N$  should compensate for jitter. It is seen that in order to control by a LB the mean rate (resp. the peak rate) of a source that has been modified by an exogeneous process, the leak rate  $R$  should be chosen as an upper-bound of the mean rate (resp. peak rate) and  $N$  should be chosen to account for the allowed "modification", i.e., the jitter.

All studies (see Rathgeb 1991 for instance) show clearly the impossibility to control any *statistical* parameter, as for instance a mean rate, or an IDI.

Two reasons can explain this: first of all, a statistical descriptor makes sense only for a source known to belong to a well-defined "class" (this is the case if the source is a Poisson process; its mean rate is described by  $\lambda$ ; but the Poisson nature of the source has to be checked at the same time. Otherwise, there exists an infinite number of ways for a process to achieve a given "mean rate"); secondly, a statistical parameter requires often lengthy measurement procedures – since they are defined as ergodic limits and thus of limited value due to the intrinsic non-stationary nature of real traffic.

However, from a practical standpoint only sample-path based schemes are of importance: no end-user would be able to predict the value to be allocated to the "mean rate" a connection is to carry, or to bound efficiently the maximum value of its IDI. Thus the most promising mechanisms are based on monitoring sample-paths which are essentially deterministic in nature.

Moreover, all studies show clearly that a single algorithm (i.e., one  $(R, N)$ -LB) can by itself control satisfactorily *only one characteristic* of the source. Also, they show the weakness of these mechanisms, where  $N$  is chosen according to the cell jitter, which is to some extent external to the source definition.

As a consequence, no attempt is made today to control anything but *operationally defined* parameters. Peak rate and jitter are defined by means of an algorithm (VSA) which gives an instantaneous answer: a cell arriving at time  $t$  is conforming or not – depending on whether or not it passes successfully the control process. This approach has gained such considerable success that the "*Sustainable Cell Rate*" has been introduced within the ATM Forum as an upper bound for the mean rate and the burst size by means of the same algorithmic procedure.

### 2.3 Quality of service

In the discussion so far the Quality of Service (QoS) has been defined in terms of delay or overflow probabilities. In almost all studies so far the preference has been on long-term or stationary overflow probabilities as a QoS measure. However, the true QoS is more correctly reflected by the rate and amount of cells that are lost. This is basically related to transient characteristics of excursions of the underlying processes. In particular in addition to the stationary overflow probability, which we denote by  $P(W > x)$  where  $W$  denotes the queue state or congestion process, the following measures associated with the overflow process need to be considered:

1. The transient overflow probability  $P(\sup_{t \leq s \leq t+T} W_s > i / W_t < x)$  which corresponds to the probability of an overflow in the next  $T$  time units given that the system is not in congestion.
2. The length of the congestion period and the number of cells or bursts arriving during a congestion period since these measures determine the amount of information lost and hence have direct implications for the AAL layer.

The above measures are related to transients and difficult to obtain analytically. These issues have been addressed recently in the context of buffered systems (modeled as  $G/GI/1$  queues) (Guillemin & Mazumdar 1995a) and unbuffered systems like multiplexers (modeled as  $M/M/\infty$  queues) in Guillemin & Mazumdar (1995b) and Iscoe *et al* (1993). One remarkable fact is that in the region of interest for QoS in ATM systems the ratio of the transient overflow probability to the stationary overflow probability can be several orders of magnitude higher. Since any overflow results in cell loss or tagging and depending on the source the durations of excursions could be large, the transient characteristics are better indicators of QoS and thus cannot be ignored. Much work needs to be done on this issue for more relevant multi-rate models which will allow the determination of the true QoS offered to a connection as seen from the AAL.

### 3. ATM Networks: the current situation

#### 3.1 Standards and their implications

ITU-T I.371 (1993) specifies what should be the traffic parameters of the source, and how to control them. Note that traffic contracts are naturally operator dependent, and as such outside the scope of Standards Organizations. However, the need to handle connections supported by more than a single network forces operators to agree on a minimum set of requirements, concerning traffic contracts and conformance with them.

One has to distinguish between *conformance testing* and *source policing*. The former refers to measurement procedures aiming at verifying if the source conforms to its negotiated traffic contract (and is clearly under the responsibility of Standards), while the latter encompasses all actions to be taken when non conforming cells are detected (and is operator specific). The network's commitment concerning QoS is thus directly related to cell conformance (Gravey & Hébuterne 1994).

ITU-T I.371 (1993) has defined the basic traffic contract at the peak cell rate of the connection. The source process is characterized by two parameters: the intrinsic Peak Rate of the connection, and the Cell Delay Variation (CDV) clumping tolerance. Specifically, for a given initial period  $T$  intrinsic to the connection ( $T$  is usually referred to as the peak emission interval of the connection (I.371 1993), cell clumping is quantified by using a reference algorithm, namely the Virtual Scheduling Algorithm (VSA) as defined in ITU-T Recommendations (I.356 1992; I.371 1993, Annex 1).

The VSA, also known as Generic Cell Rate Algorithm (GCRA) within the ATM Forum (1993), depends on two parameters, namely the increment  $I$  and the limit  $L$ . The increment  $I$  is usually equal to some reference period (e.g., the peak emission interval of a connection) and the limit  $L$  is some tolerance around this period, accounting for random cell transfer delays. In other words, when a cell stream is observed with a VSA( $I, L$ ) to check conformance with respect to the cell rate  $1/I$ , a certain number of cells, up to some limit related to  $L$ , are allowed to arrive with an instantaneous cell rate larger than  $1/I$ . Specifically, a group of cells forms a cell clump when the VSA( $T, \tau$ ) is ahead of the current time and the parameter  $\tau$  quantifies the cell clumping effect on a connection in the strict sense (resp. the wide sense with a precision  $\varepsilon$ ) if all cells (resp. a fraction of at least  $(1 - \varepsilon)$  of cells) pass the VSA( $T, \tau$ ). The impact of cell clumping on the peak cell rate control

function located in the Usage/Network Parameter Control (UPC/NPC) and then on resource allocation has been investigated in Roberts & Guillemin (1992) and Boyer *et al* (1992) respectively.

The first public ATM networks offer subscribers only peak cell rate traffic contracts. Note that it amounts to CBR source characterization only. However, the presence of jitter forces a negotiation based on the couple  $(T, \tau)$  at the call set up.

The specification of cell dispersion within the standardization bodies is not as complete as that of cell clumping, especially because its close relationship with AAL issues.

Concerning policing actions, I.371 describes two possibilities: either the non-conforming cells are immediately discarded, or they are marked as non-conforming (*tagging*). A tagged cell is allowed to enter the network, but when congestion is detected within the network, tagged cells are destroyed first. Tagging uses the *Cell Loss Priority* bit in the header. This option implies that efficient selective discard mechanisms are implemented in the network (see below and e.g. Gravey & Hébuterne (1991), Gravey *et al* (1991), Kroner *et al* (1991) for more details).

The CLP bit allows another option, namely to offer two levels of QoS, more precisely two levels of Cell Loss Rate (CLR). The corresponding traffic contract implies to control both the CLP=0 flow and the aggregated CLP=0+1 flow. Serious issues have already been raised concerning the control of such hybrid contracts (Gravey *et al* 1991; Gravey 1994; Guillemin *et al* 1995). For instance, the control on the aggregated 0+1 flow could "destroy" (or mark as non-conforming) CLP=0 cells already accepted by the specific control.

*Connection Acceptance Control* procedures (CAC) have to take into account the jittered stream, in order to allocate resources to accepted connections. This is done by assuming that all sources behave according to the worst-case the control tolerates, for safety. The network utilization is thus the consequence of the following trade-off:

- either the connections request a low value for the CDV tolerance (the jitter the flow is allowed to be affected by) – say, less than 1 ms; this clearly limits the ATM gains, and increases the cost of subscribers equipments (need to shape outgoing traffic);
- or the connection may have a relatively high CDV tolerance (i.e., several ms) – i.e., the network accepts high jitter, the control device having accordingly unfavourable worst-case sources. This results in quite poor network utilization.

It is however possible to overcome this dilemma, by performing *active* control actions: Instead of using the negotiated parameters  $(T, \tau)$  for conformance testing only, it has been proposed to use them to reconstruct the initial stream, as far as possible. This is a "reshaping" action: arriving cells are buffered and sent with a spacing at least equal to  $T$ . A control action is performed at the same time by limiting the buffer size, so as to discard very long bursts (possibly due to excessive jitter or to malevolent sources).

The "Spacer-Controller" (Boyer *et al* 1992) has a great advantage as compared with a "passive" control, like Leaky Buckets. Instead of limiting the size of bursts entering the network, it suppresses them, the outgoing flow is strictly conforming to its peak rate (actually, a residual jitter may be observed, due to multiplexing inside the Spacer Controller).

As a consequence, CAC procedures in a network which make use of the Spacer-Controller may be implemented without any reference to worst-case traffic. Note however that the choice to use such a shaping mechanism is operator specific.

The ATM-Forum (1993) extends ITU-T Recommendations, especially by defining the *Sustainable Cell Rate* Traffic Descriptor. It has been introduced by the ATM Forum to characterize *within the ATM Layer* and by means of an algorithmic rule, the statistical activity on an ATM connection. It is composed of two traffic parameters, namely the Sustainable Cell Rate  $1/T_{SCR}$  and the Intrinsic Burst Tolerance (IBT)  $\tau_{IBT}$ . Roughly speaking, the SCR is an upper bound on the source mean cell rate, while the IBT sets a limit on the size of bursts at peak cell rate. The definition of the SCR traffic descriptor relies on an intuitive approach aimed at defining an envelope for the traffic generated by an ATM source. The algorithmic rule used to define the SCR traffic descriptor is the GCRA, which has been adopted within the standardization bodies as the reference algorithm for cell rate monitoring (I.371 1993) and cell conformance testing (I.356 1992; Guillemin et al 1994a).

As for "0,0+1" contracts, interworking of Peak Rate and SCR controls on a given connection have to be carefully defined. In recent work Guillemin et al (1994b) showed that the specification of the SCR traffic descriptor, as stated by the ATM Forum, is incomplete. As it is defined now, the SCR characterization leaves the source with an infinite number of  $(T_{SCR}, \tau_{IBT})$  couples to choose from. This ambiguity is likely to be a potential source of problems. While admissible couples could be considered equally suitable from the source perspective, such is not the case from a networking point of view. The parameters of the SCR determine the worst case traffic pattern allowed in the network, which is generally recognized as the basis for resource allocation. A poor choice of SCR might thus involve lower network utilization (Rosenberg and Hébuterne (1994)); furthermore, the consequences depend on the multiplexing scheme and other network specific factors. In all cases, further information should be provided in order to insure that values for  $1/T_{SCR}$  and  $\tau_{IBT}$  are selected properly (Guillemin et al 1994b). It has been shown that in order to obtain a useful notion of the SCR from these different points of view it is necessary to define a "cost functional" and the SCR which in some sense optimizes this functional for "worst-case" traffic can lead to a useful measure.

The ITU-T Recommendations as well as the ATM-Forum (1993) mention other traffic control mechanisms which are discussed below in the section dealing with the "future" of ATM networks. One of them is the *Explicit Forward Congestion Indication* (EFCI), which is viewed as a "Traffic Control Function" in the ATM-Forum, while it is a "Congestion Control Function" within ITU-T (I.371 1993). This mechanism can be considered a high-level feedback or reactive mechanism which is only used in the case when congestion is present. Another one is termed with the generic name of *Fast Reservation Mode*.

Before dealing with the "future" of ATM networks (in terms of what remains to be done to provide an efficient broadband integrated service network), we describe below some current European experimental ATM networks intended for testing on a large scale a certain number of ideas and mechanisms.

### 3.2 Examples: the ATM-Pilot, the Brehat Project

European Public Network Operators have devised a public ATM network within the ATM-Pilot project. The first "public" ATM connection ought to be opened by the

end of 1994 (ATM-Pilot 1994).

The options taken during the specification phase of the network reflect the caution regarding the possibility of elaborate controls. Connections are offered on the basis of peak cell rate traffic contracts. No attempt is made to make use of the CLP bit to carry two levels of loss-sensitive cells. Concerning the control mechanisms and their actions, there are naturally operator specific, and will be defined soon.

However, the network will serve as a first step towards more efficient multiplexing schemes. The goal is to gain insight with ATM traffic and multiplexing mechanisms through a few benchmark services and experimentations.

"Brehat" is another experimental ATM network, led by France-Telecom (see Onno *et al* 1993). The first step of the project (December 1993) offers peak rate connections and control by means of Leaky Buckets. The second phase, expected for the end of 1994, will make use of the Space Controller (Boyer *et al* 1992) for the control of ingoing traffic.

#### 4. ATM Networks: the future

The previous Section has presented the current situation on public ATM networks (at least in the European context). No attempt is made in this first step to offer services beyond the peak rate connections.

Evidently, this is to be seen as a first step, necessary to get acquainted with traffic phenomena, and with source and control schemes. The next step will consist in increasing network efficiency by offering some kind of multiplexing. The ultimate step would be to introduce multiplexing schemes allowing the network to be operated at utilization rates around 80% – and sources multiplexed according to their mean rates. This is not the goal of this paper to justify this evolution or to develop the related technical points. Only, for such multiplexing schemes to be implemented safely, the corresponding traffic control mechanisms have to be devised.

At this point, it is worth emphasizing a few remarks.

First of all, the question arises of the need for such improvements *within the ATM network*. That is, for all additional functions beyond the basic ones, the decisions to be taken are: should the function take place in the ATM network; or should it be integrated inside one (or several) AAL; or should a specific overlay network be built over the AAL, dedicated to specific service, and implementing the function. This is typically true for error correction, but could also be discussed for multiplexing. In the case where multiplexing schemes would be in use in AALs or in dedicated networks, the need of traffic control functions would not vanish: they would shift to these levels automatically.

Note however that different solutions may coexist at different levels, users and network managers being able to take advantage of the best one according to the application. For instance, reactive mechanisms (see below) behave quite differently if operated in a large public network (where round trip delays may reach tens of milliseconds) and in a Local Area Network (LAN) of limited geographic extension.

##### 4.1 Statistical multiplexing schemes

The effectiveness of ATM networks as efficient carriers of information is going to depend very heavily on the efficiency of multiplexing. The ideal multiplexing mechanism – at least for the users' viewpoint, consists in sources sending their traffic

blindly, once the connection is set-up. This is usually referred to as *straightforward statistical multiplexing*. However, simple multiplexing where all the users have access to the available bandwidth will not work efficiently enough and more elaborate schemes based on bandwidth reservation schemes will be needed. Various mechanisms can be envisaged based on movable windows, thresholds or partial bandwidth sharing schemes. One of the most promising schemes in the ATM context is the so-called *generalized processor sharing* (GPS) or weighted fair queueing scheme (WFQ).

*Weighted Fair Queueing* (WFQ) is a service discipline designed to provide both delay and throughput performance guarantees, under the assumption that connections are constrained by LBs. This scheme was introduced in Parekh & Gallagher (1993, 1994) as a passive means of controlling congestion in high speed integrated services networks. It is, in essence, a generalization of round robin service discipline, where each session is guaranteed a minimal throughput and any remaining server capacity is split proportionally among connections.

All the performance bounds obtained for such schemes are essentially deterministic based on the assumption that the traffic is  $(\sigma, \rho)$  constrained. What is needed are analyses based on statistical assumptions. These models fall under the class of the so-called multi-rate loss models and these models both under straight forward multiplexing and bandwidth reservation have been addressed by Kelly (1991) and Theberge & Mazumdar (1994, 1995).

Another promising scheme is based on the "*Fast Reservation Mode*" (FRM) which covers all protocols by which the resources allocated to a connection can be modified in real time during the duration of the call.

A proposal for such a mechanism is made by Tranchier *et al* (1992). The connection is set-up based upon the usual peak rate contract. The protocol is operated by a command unit in the UPC/NPC. Upon request of the source, the protocol sends a Request for Modification along the path of the connection. At each node, the Request is processed, and either accepted or rejected. If it is accepted, the resource allocated to the connection is updated. After completing its travel, the Request returns the command unit. If the bandwidth increase is accepted, the command updates the parameters  $(T, \tau)$  of the traffic control device (LB, VSA) in the UPC.

The advantage of this multiplexing mechanism is to permit the network to accept long bursts, without long buffers. Note that the protocol may be operated by the network or by the user (within a VP, e.g.). See Tranchier *et al* (1992) for additional discussion.

For such schemes to guarantee the negotiated QoS, the acceptance control must be carefully tuned.

Assume the peak rates are already controlled – and preferably enforced. For simplicity let us assume identical sources, with peak rate  $D$  and mean rate  $m$ . Let  $C$  be the link capacity. With peak rate allocation, the number of simultaneous connections is around  $C/D$  (assuming 100% efficiency). Statistical multiplexing allows a number of connections between  $C/D$  and  $C/m$ .

Statistical multiplexing requires us to define traffic parameters characterizing the statistical behaviour of the source (i.e., characterization beyond the peak cell rate). For this purpose, the ATM-Forum (1993) has defined the Sustainable Cell Rate. This descriptor limits the size of the bursts the source is allowed to send at the peak rate of the connection, upperbounding the actual mean rate. It is defined with respect to a VSA.

From the viewpoint of traffic control, FRM can be seen as statistical multiplexing using CBR sources – i.e., the only traffic control operates on the usual parameters ( $T, \tau$ ) of the peak rate management – with the improvement that these parameters are dynamically controlled.

## 4.2 Explicit notification

The principle of reactive, explicit congestion notification is well known. It has been adopted for instance in the Frame Relay networks. See e.g. Makrucki (1991) and Newman (1993) for its use in the ATM context. Two possibilities may be envisaged:

- Forward Congestion Notification (FCN). Once a node inside the network detects a congestion, it informs the receiver; the receiving end has then to take any flow control action to slow down the source.
- Backward Congestion Notification (BCN). In this case, the network element informs directly the sender (in the “backward” direction, implying a bidirectional connection).

The use of the FCN mechanism is mentioned, both in ITU-T (I.371 1993) and in the ATM-Forum (1993). Since an ATM connection is bidirectional, backward notification is always possible, *via* FCN. However, since responding to FCN/BCN is optional for end users, such mechanisms cannot be used alone.

FCN/BCN could offer interesting traffic management opportunities. This is especially true in local area network. Consider the typical LAN, with round trip delays around 50 – 100  $\mu$ s. If the duration of congestions is much larger than this time, the source may usefully react. Note that 100  $\mu$ s is the time needed to send around 35 cells on a 150 Mbit/s link. In Wide Area Networks, or in large public networks, round trip delays may reach tens of milliseconds. In this case, the source could have sent up to  $10^4$  cells before receiving any notification. Note too that the LAN environment may be considered as more “cooperative” as compared with a public network. See Newman (1993) for a more complete discussion.

Finally, FCN/BCN mechanisms may help efficiently to regulate traffic and recover from congestion, especially in a LAN environment. As traffic control mechanisms, it remains to integrate them inside a complete set (including enforcing mechanisms); it remains also to determine which traffics would benefit from such reactive mechanisms.

## 5. Conclusions

In this paper, we have surveyed a number of topics related with the control of traffic in ATM networks. The need of efficient control schemes comes from the flexibility ATM is to offer to its users, in order to achieve a true multiservices network.

It was soon understood that it was both useless and impossible to control statistical characteristics of traffic sources. So the studies on traffic control have focussed on defining operational parameters related with traffic phenomena. These parameters are negotiated at call set-up, and then monitored during the connection duration through deterministic and real-time algorithms. Recall, that for a traffic parameter to be called “operational”, a measurement algorithm (or a conformance test) has to be defined with it.

Until now, Peak Cell Rate is defined this way (in duality with Cell Delay Variation), and the first implementations of public ATM networks should offer only peak rate based connections.

However, the need of more efficient multiplexing schemes spurs traffic specialists to devise new control mechanisms. Any multiplexing scheme is based on some traffic characteristics of the source. This must correspond to an operational parameter, and to Connection Acceptance Control procedures.

Statistical multiplexing (based on the Sustainable Cell Rate), Fast Reservation schemes, operating at the end user data units timescale, are declared candidates for these extensions. However, it is unclear if the ATM network is the preferred layer for elaborate schemes, or if elaborate multiplexing has to be offered on a per service basis, that is inside the ATM Application Layer.

In that case, the new traffic control algorithms would have to migrate towards upper layers (AALs or above). Nevertheless, traffic control will remain an important task of B-ISDN traffic management.

## References

- Anantharam V, Konstantopoulos T 1994 Reflection mappings and flow control. Preprint, University of Texas, Austin (presented at Workshop on Stochastic Models, Ulm, Germany)
- ATM Forum 1993 The ATM Forum: UNI Specification. Version 3.0, Boston
- ATM-Pilot 1994 Info-Booklet of the ATM-Pilot Network
- Berger A E 1994 Desirable properties of traffic descriptors for ATM connections in broadband ISDN. *Proceedings ITC 14* (eds) J Labetoulle, J W Roberts (New York: Elsevier Science)
- Boyer P, Guillemin F, Serel M, Coudreuse J P 1992 Spacing cells protects and enhances utilization of ATM network links. *IEEE Networks*
- CNET 1991 Special issue on ATM. *Echo des Recherches* (English issue) France Telecom CNET
- Cruz R L 1991a A calculus for network delay, part I: Network elements in isolation. *IEEE Trans. Info. Theory* 35: 114-131
- Cruz R L 1991b A calculus for network delay, part II: network analysis. *IEEE Trans. Info. Theory* 35: 131-141
- Di Vecinia G, Courcoubetis C, Walrand J 1994 Decoupling bandwidths in queues. Preprint, University of California, Berkeley
- Doshi B T 1993 Deterministic rule based traffic descriptors for a broadband ISDN: worst case behavior and connection acceptance control. *Proceedings of Globecom'93 Houston*, paper 48.A.1

Elwalid A, Mitra D 1993 Effective bandwidths of general Markovian sources and admission control of high speed networks. *IEEE/ACM Trans. Networking* 1: 329-343



- Gravey A 1994 Cell conformance and quality of service guarantees in ATM networks. *14th International Teletraffic Congress Antibes* (France)
- Gravey A, Hébuterne G 1991 Mixing time and loss priorities in a single server queue. *13th International Teletraffic Congress Stockholm*
- Gravey A, Hébuterne G 1994 Cell loss ratio commitments in ATM networks (to be presented in ICC'94)
- Gravey A, Boyer P, Hébuterne G 1991 Tagging versus strict rate enforcement in ATM networks. *Globecom '91 Phoenix*
- Guillemin F, Mazumdar R 1995 On excursions of the workload process in  $G|GI|1$  queues. *Stochastic processes and their applications* (to appear)
- Guillemin F, Levert C, Rosenberg C 1994a Cell conformance testing with respect to the peak cell rate in ATM networks. *Comput. Networks ISDN Syst.* (to appear)
- Guillemin F, Rosenberg C, Mignault J 1994b Guidelines for sustainable cell rate specification in ATM networks. *Comput. Networks ISDN Syst.* (submitted)
- Guillemin F, Mazumdar R, Simonian A 1995 On heavy traffic approximations for transient characteristics of  $M|M|\infty$  queues. *J. Appl. Probab.* (to appear)
- Hui J Y 1988 Resource allocation in broadband networks. *IEEE Selected Areas Commun.* JSAC-6: 1598-1608
- I.121 1989 ITU-T Recommendation I.121: *Broadband aspects of ISDN*. Blue Book, vol. III.7, Geneva
- I.356 1992 ITU-T Recommendation I.356: *B-ISDN ATM layer cell transfer performance*, Geneva
- I.371 1993 ITU-T Draft Recommendation I.371: *Traffic control and congestion control in B-ISDN*. Geneva
- Iscoe I, McDonald D, Qian K 1993 Capacity of ATM switches. *Ann. Appl. Probab.* 3: 277-295
- Kelly F P 1991a Loss networks. *Ann. Appl. Probab.* 1: 319-378
- Kelly F P 1991b Effective bandwidths at multi-class queues. *Queueing Syst.* 9: 5-15
- Kroner H, Boyer P, Gravey A, Hébuterne G 1991 Priority management in ATM switching. *IEEE J. Selected Areas Commun.* : 418-427
- Labetoulle J, Roberts J W (eds) 1994 *Proceedings of the International Teletraffic Congress, ITC 14* (New York: Elsevier Science)
- Leland W E, Taqqu M S, Willinger W, Wilson W S 1994 On the self-similar nature of ethernet traffic. *IEEE/ACM Trans. Networking*

- Makrucki B A 1991 On the performance of submitting excess traffic to ATM networks. *IEEE Globecom'91* Phoenix
- Minzer S E 1989 Broadband ISDN and asynchronous transfer mode (ATM). *IEEE Commun. Mag.*
- Newman P 1991 Backward explicit congestion notification for ATM local area networks. *IEEE Globecom'93* Houston
- Norros I 1993 Studies on a model for connectionless traffic based on fractional Brownian motion. *Presented at the Conf. on Appl. Probab.* Paris (to appear in *Queueing Syst.*)
- Onno G, Lemonier M, Kerbérenes G, Gonzales J 1993 Services and applications in the Brehat Project. *ISSLS'93* Vancouver
- Parekh A, Gallager R 1993 A generalized processor sharing approach to flow control intergrated services networks - the single node case. In *Proc. Infocom'92* Florence. Also 1993 in *IEEE/ACM Trans. Networking* 1
- Parekh A, Gallager R 1994 A generalized processor sharing approach to flow control intergrated services networks - the multiple node case. *IEEE/ACM Trans. Networking* 2: 137-150
- Rathgeb E 1991 Modeling and performance analysis of policing mechanisms for ATM networks. *IEEE J. Selected Areas Commun.* 9:
- Roberts J W (ed.) 1991 COST 224, Performance evaluation and design of multi-service networks. Final report, Commission of the European Communities
- Roberts J W, Guillemin F 1992 Jitter in ATM networks and its impact on peak cell rate control. *Perform. Eval.* Special Issue on Modelling of High Speed Telecommunications Systems
- Rosenberg C, Hébuterne G 1994 Dimensioning traffic control devices in an ATM network. *Proceedings Conference on Telecommunication Systems* Nashville
- Rosenberg C, Lague B 1994 A heuristic framework for source policing in ATM networks. *IEEE/ACM Trans. Networking* (to appear)
- Rosenberg C, Guillemin F, Mazumdar R 1993 On quantile measures for traffic characterization in B-ISDN. *Proc. ITC Specialists Seminar* Bangalore
- Theberge F, Mazumdar R 1994 Approximations for loss probabilities in large multi-rate systems. Preprint INRS-Telecommunications, (also submitted for publication)
- Theberge F, Mazumdar R 1995 Multi-rate loss systems with bandwidth reservation: Algorithms and analysis. *IEEE Selected Areas Commun.* Special Issue on Foundations of Networking (to appear)
- Tranchier D, Boyer P, Rouaud Y, Mazeas J Y 1992 Fast bandwidth allocation in ATM networks. *ISS'92* Yokohama City

Turner J 1986 New directions in communications (or which way in the information age ?). *Proceedings of the Zurich Seminar on Digital Communications* Zurich, pp 25-32

Whitt W 1995 Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues. *Telecommun. Syst.* (to appear)



# PVU and wave-particle splitting schemes for Euler equations of gas dynamics

S M DESHPANDE, N BALAKRISHNAN and S V RAGHURAMA RAO

CFD Laboratory, Department of Aerospace Engineering, Indian Institute of Science, Bangalore 560 012, India

**Abstract.** A new way of flux-splitting, termed as the wave-particle splitting is presented for developing upwind methods for solving Euler equations of gas dynamics. Based on this splitting, two new upwind methods termed as Acoustic Flux Vector Splitting (AFVS) and Acoustic Flux Difference Splitting (AFDS) methods are developed. A new Boltzmann scheme, which closely resembles the wave-particle splitting, is developed using the kinetic theory of gases. This method, termed as Peculiar Velocity based Upwind (PVU) method, uses the concept of peculiar velocity for upwinding. A special feature of all these methods is that the unidirectional and multidirectional parts of the flux vector are treated separately. Extensive computations done using these schemes demonstrate the soundness of the ideas.

**Keywords.** Upwind methods for Euler equations; wave-particle splitting; Boltzmann schemes; peculiar velocity based upwinding.

## 1. Introduction

Computational Fluid Dynamics (CFD) a new revolutionary tool in aerodynamic design and analysis, is progressing at a very fast pace and enormous developments in algorithm development, grid generation and postprocessing have taken place in the last twenty years. CFD aims at solving numerically the partial differential equations of fluid dynamics and hence requires the use of digital computers which have progressed over years at a breathtaking pace. The ultimate aim is to obtain the direct numerical simulation of unsteady Navier-Stokes equation for an arbitrarily large Reynolds number for practical geometrical configurations such as aircraft, missile, launch vehicle, helicopters, ships or submarines. This aim has not been realised as yet in spite of the development of massively parallel computers and numerical solution of only some simple flow problems such as flow through a channel, flow past a flat plate, flow around a cylinder have been attempted using Direct Numerical Simulation of Turbulence (DNS). Even though it is true that the numerical solution of Navier-Stokes equation must be sought for every flow problem, there are many

problems for which aerodynamic analysis and design can be done using lower level approximations which include potential flows governed by Laplace's or full potential equation, attached flows for which Euler equations are a valid approximation, thin layer Navier-Stokes equation (TLNS) and so on. In this paper we are going to concentrate on numerical methods for the solution of Euler equations of gas dynamics. Many problems arising in the aerodynamic design and analysis of the aerospace vehicles require the solution of Euler equations. Some notable examples are flow past delta wings, flow through ramjet type intakes, external supersonic flow past launch vehicles, hypersonic reentry flows. These equations also need to be solved in cases of flow problems where viscous effects are approximately taken into account by solving the well-known boundary layer equations which again require the inviscid solutions as the input. Further a good Euler solver is a prerequisite to the development of a good Navier-Stokes solver. Therefore constructing numerical schemes for solving the Euler equations has been one of the principal subjects of research among the CFD community for the last decade.

An ideal numerical scheme, none exists at all as of now, is the one which satisfies several requirements:

- (1) It must be robust, that is, it should work for a wide variety of flow conditions covering all geometrical shapes encountered in practice with Mach number and angle of attack varying over a large range and various boundary conditions. Further, the code based on this method must work over different type of grids such as unstructured tetrahedral, structured hexahedral, clustered/unclustered and hybrid grids with the grid aspect ratio and the skewness varying widely.
- (2) The numerical schemes must be accurate enough to capture shocks, contact surfaces, shear layers with acceptable accuracy and resolution.
- (3) The scheme must be easy to code and be computationally inexpensive and must rapidly converge to the steady state whenever necessary.
- (4) It must be adaptable to computer architecture in the sense that it should be vectorisable and should admit massive parallelism.

In summary, an ideal numerical method must be *robust, efficient, accurate, rapidly converging* and *adaptable to new emerging computer architecture*. Unfortunately, in spite of several years of research in algorithm development, such an ideal scheme does not exist till date. The search for the elusive best scheme is on.

Several interesting characteristics have emerged about numerical schemes after several years of intensive research. First, it has been found that even if a numerical scheme is consistent and stable in the von Neumann sense it need not converge to the solution. Lax-Wendroff scheme and MacCormack scheme do require Total Variation Diminishing (TVD) fixes in order to suppress pre-shock and post-shock wiggles encountered in capturing discontinuities. Without the TVD fix these and many other second order accurate schemes can cause violent oscillations in the flow variables leading to negative values of pressure and density. The mathematical theory of stability of numerical schemes for nonlinear partial differential equations together with a boundary condition treatment is just not available today to designers of numerical schemes to serve as guidelines for algorithm development. Secondly many first-order schemes possessing TVD property do not have solution reliability. For example Roe's method (Roe 1981) based on the approximate Riemann solver can admit unphysical shocks (called carbuncle shocks), may fail in capturing large rarefaction waves (Quirk 1992) and is known to converge to rotationally asymmetric solutions

in case of supersonic flow past a hemisphere (P K Sinha, private communication). A lack of robustness has also been reported for the Osher method (Osher & Solomon 1982) in capturing strong detached shock. Many fixes have been proposed to cure some of these failures, but these fixes are not universal and are known to destroy solution accuracy. Harten's entropy fix is known to spoil the high resolution property of Roe's scheme (Quirk 1992). Flux Vector Splitting methods on the other hand do possess solution reliability and capture shocks and large rarefaction waves without any problem but they are notorious for smearing the contact discontinuity (van Leer 1990). They also cause unacceptably large smearing of boundary layer (van Leer 1990). Central differencing schemes with artificial dissipation (Jameson *et al.* 1981) have their own problems. These schemes contain many tuning parameters which must be adjusted for robustness, accuracy and convergence. It has now become clear that having the correct amount of dissipation is the key to the design of an ideal numerical scheme. MacCormack (1990) has rightly observed *it is all dissipation*. Lastly, it has been found that a numerical scheme that solves a discrete mathematical model, which is an approximation to the partial differential equation purported to be solved, has adequate robustness if the discrete mathematical model mimics the physics of the flow as closely as possible. As an example to illustrate this principle we may cite the case of upwind methods which take into account the signal propagation property of the Euler equations by appropriately choosing the stencil of grid points. The Flux Vector Splitting schemes as noted above are very robust. One of the Flux Vector Splitting Schemes due to Deshpande & Mandal (Deshpande 1986c, Mandal 1989, Mandal & Deshpande 1993) called the Kinetic Flux Vector Splitting scheme (KFVS) has resulted in an extremely robust code BHEEMA routinely used at DRDL, Hyderabad for computing low speed, high speed and hypersonic flows around a variety of practical configurations. In the absence of clearcut and complete set of guiding principles to be used while constructing an ideal numerical method, we follow the following methodology while searching for new algorithms, (1) New directions and lines of research leading to novel algorithm development must be constantly explored. (2) Once these ideas take a concrete shape in the form of a scheme then it should be tested for robustness, accuracy, solution reliability, convergence property and adaptability by trying it on the largest set of fluid flow problems possible. (3) Suitable modifications in the algorithm should be progressively introduced for slowly but steadily marching towards the best scheme. We study in detail a new line of research called wave-particle splitting which is closely related to an allied idea termed Peculiar Velocity based Upwind (PVU) methods at the Boltzmann level.

## 2. Wave-particle splitting

### 2.1 Acoustic flux vector splitting

The Euler equations of gas dynamics can be written in the differential form of the conservation law as,

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} + \frac{\partial H}{\partial z} = 0, \quad (1)$$

where  $U$  = Conserved variable vector,  $\mathbf{Q} = i_x F + i_y G + i_z H$  = flux vector and are defined by the relations,

$$U = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ e \end{bmatrix}; F = \begin{bmatrix} \rho u \\ p + \rho u^2 \\ \rho uv \\ \rho uw \\ (e + p)u \end{bmatrix}; G = \begin{bmatrix} \rho v \\ \rho vu \\ p + \rho v^2 \\ \rho vw \\ (e + p)v \end{bmatrix}; H = \begin{bmatrix} \rho w \\ \rho wv \\ \rho vw \\ p + \rho w^2 \\ (e + p)w \end{bmatrix}. \quad (2)$$

Here  $\rho$  = mass density,  $u, v, w$  are the components of velocity along the  $x, y, z$  directions,  $p$  = pressure and  $e$  = specific total energy per unit volume and is given by,

$$e = \frac{p}{(\gamma - 1)} + \frac{1}{2}\rho(u^2 + v^2 + w^2). \quad (3)$$

Even though in (2)  $F, G, H$  are expressed as functions of the primitive variables  $\rho, u, v, w, p$  they can also be expressed as functions of the conserved variables  $U$ . Equations (1) are nonlinear hyperbolic partial differential equations for the unknown  $U$  and must be solved with suitable boundary conditions. Because of hyperbolicity these vector conservation laws involve propagation of waves which are the well-known examples of nonlinear waves. Any numerical scheme if it claims to mimic the physics of the flow, must take into account the appropriate directions of information propagation. Currently there are two different ways of incorporating hyperbolicity into numerical schemes. The first one is the *flux difference splitting* where Riemann's initial value problem is solved approximately. Two important flux difference splitting schemes are currently available, one by Roe (1981) and the other by Osher (Osher & Solomon 1982). The underlying physical model in these approaches is the interaction among cells through wave propagation. The variables are assumed to be constant within a cell and therefore fluid variables undergo jumps when we cross a cell face. These jumps are broken into waves using the Riemann solver and thus the cells interact with one another through waves. The second approach involves splitting of the flux vector into two parts  $F^+, F^-$  (for the  $x$ -component of  $\mathbf{Q}$ ) such that the Jacobian  $\partial F^+ / \partial U$  has all positive eigenvalues and the Jacobian  $\partial F^- / \partial U$  has all negative eigenvalues. The flux vector splitting of Steger & Warming (1981), van Leer (1982) fall under this class. The underlying physical model here is the particle model. The flux vector splitting schemes can be regarded as a generalisation of the Courant-Isacson-Rees scheme or as a Beam scheme (Sanders & Prendergast 1974) wherein it is tacitly assumed that there are in case of a 1-D problem two beams of particles moving in opposite directions. For this reason the FVS is also sometimes referred to as pseudo-particle method. The particle nature of FVS is even more obvious when we consider the KFVS method which is derived from the Boltzmann equation of kinetic theory of gases using the moment method strategy (Deshpande 1986c). An interesting question arises here as to whether we can construct new schemes using both the particle model and the wave model of fluid flow. There is a strong physical basis for seeking such a wave-particle model. For, consider 1-D



Euler equations in primitive variables,

$$\left. \begin{aligned} \frac{\partial \rho}{\partial t} + u \frac{\partial \rho}{\partial x} + \rho \frac{\partial u}{\partial x} &= 0 \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{1}{\rho} \frac{\partial p}{\partial x} &= 0 \\ \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + \gamma p \frac{\partial u}{\partial x} &= 0 \end{aligned} \right\}. \quad (4)$$

We have the famous convective derivative appearing in all the three equations and its appearance is due to the motion of fluid elements along particle paths. In order to demonstrate the wave nature let us write (4) in the matrix form.

$$\frac{\partial V}{\partial t} + \tilde{A}^t \frac{\partial V}{\partial x} + \tilde{A}^a \frac{\partial V}{\partial x} = 0, \quad (5)$$

where,

$$V = \begin{bmatrix} \rho \\ u \\ p \end{bmatrix}; \tilde{A}^t = \begin{bmatrix} u & 0 & 0 \\ 0 & u & 0 \\ 0 & 0 & u \end{bmatrix}; \tilde{A}^a = \begin{bmatrix} 0 & \rho & 0 \\ 0 & 0 & 1/\rho \\ 0 & \gamma p & 0 \end{bmatrix}. \quad (6)$$

The eigenvalues of  $\tilde{A}^a$  can be easily shown to be

$$\lambda(\tilde{A}^a) = 0, \pm a \quad (7)$$

where  $a = (\gamma p / \rho)^{1/2} = \text{sonic speed}$ . Thus the third term in (5) can be considered to represent propagation of acoustic waves in both directions in case of 1-D and all directions in case of multidimensional flows. We therefore take the view that the motion of fluid is a mixture of particle behaviour (the basis of Lagrangian description) and wave behaviour. This view is also supported by the fact that the rate of change of a conserved quantity (momentum or energy) in a control volume is due to the change caused by the transport of fluid (particle like) and the change caused by the pressure acting on the control volume. We can also regard the dynamics of fluid as consisting of the advection of the fluid element along particle paths and further these fluid elements are subjected to dilatation or contraction thus sending pressure waves into the domain. It is therefore physically meaningful to regard the fluid motion as partly particle-like and partly wave-like.

We have therefore a natural splitting of the flux vector into transport part and acoustic part and hence we write the 1-D Euler equation in the form,

$$\frac{\partial U}{\partial t} + \frac{\partial F^t}{\partial x} + \frac{\partial F^a}{\partial x} = 0, \quad (8)$$

where,

$$F^t = \begin{bmatrix} \rho u \\ \rho u^2 \\ e u \end{bmatrix}; F^a = \begin{bmatrix} 0 \\ p \\ p u \end{bmatrix} \quad (9)$$

The corresponding flux Jacobians  $A^t = \partial F^t / \partial U$  and  $A^a = \partial F^a / \partial U$  are given by,

$$A^t = \begin{bmatrix} 0 & 1 & 0 \\ -u^2 & 2u & 0 \\ e u & e & u \end{bmatrix}, \quad (10)$$

$$A^a = \begin{bmatrix} 0 & 0 & 0 \\ \frac{(\gamma-1)}{2}u^2 & -(\gamma-1)u & (\gamma-1) \\ -\frac{pu}{\rho} + \frac{(\gamma-1)}{2}u^3 & \frac{p}{\rho} - (\gamma-1)u^2 & (\gamma-1)u \end{bmatrix}. \quad (11)$$

The eigenvalues of the matrices are given by

$$\lambda(A^t) = u, u, u; \lambda(A^a) = 0, \pm[(\gamma-1)/\gamma]^{\frac{1}{2}} a. \quad (12)$$

Observe that the corresponding matrices  $\tilde{A}^t$  and  $\tilde{A}^a$  in the primitive variable representation have the eigenvalues  $\lambda(\tilde{A}^t) = u, u, u$  and  $\lambda(\tilde{A}^a) = 0, \pm a$ . Thus there is difference between  $\lambda(\tilde{A}^a)$  and  $\lambda(A^a)$  because of the transformation from  $V$ -representation to  $U$ -representation.

Now let us look at the construction of an upwind scheme for the Euler equations based on the wave-particle splitting idea. The term involving the flux  $F^t$  does not pose any problem as far as implementing the upwinding is concerned. Based on whether  $u$  is positive or negative  $\partial F^t/\partial x$  can be backward or forward differenced for enforcing upwinding. The task of enforcing upwinding for the term  $\partial F^a/\partial x$  is slightly more complex due to the mixed eigenvalues of  $A^a$ , that is eigenvalues are of mixed sign. Many variations of similar flux splitting exist but they are different from our wave-particle splitting. The Convective Upwind and Split Pressure (CUSP) schemes of Jameson (1993) split the flux as,

$$F^c = \begin{bmatrix} \rho u \\ \rho u^2 \\ H u \end{bmatrix}; F^p = \begin{bmatrix} 0 \\ p \\ 0 \end{bmatrix}, \quad (13)$$

where,  $H = e + p$ . These are different from  $F^t$  and  $F^a$  given by (9). Since the eigenvalues of  $\partial F^c/\partial U$  are  $u, u$  and  $\gamma u$  while those of  $\partial F^p/\partial U$  are  $0, 0$  and  $-(\gamma-1)u$ , a splitting with,

$$F^+ = F^c, F^- = F^p, \quad (14)$$

leads to stable scheme as done by Denton (1983). As remarked by Jameson this scheme does not reflect the true zone of dependence in supersonic flow. He then modifies the scheme by an appropriate choice of diffusive flux. Again the Advective Upstream Splitting Method [AUSM] of Liou & Steffen (1991) as different from the present wave-particle splitting in the sense that  $F^t$  and  $F^p$  chosen by them are not the same as the transport and acoustic fluxes given by (9). Conceptually we are using a physically meaningful basis for splitting  $F$  into  $F^t$  and  $F^a$ .

We (Balakrishnan & Deshpande 1991, 1992a, 1994a, 1994b, Balakrishnan and Raghurama Rao 1992b) have further split  $A^a$  following a method similar to that of Steger & Warming (1981). Here we apply this method to  $F^a$  as against Steger & Warming (1981) who considered the total flux  $F$  instead. Towards this end we require the canonical form

$$A^a = R^a \Lambda^a (R^a)^{-1}, \quad (15)$$

where  $R^a$  = matrix of right eigenvectors of  $A^a$  and  $\Lambda^a$  = diagonal matrix and these are given by

$$R^a = \begin{bmatrix} 0 & 1 & 0 \\ 1 & u & 1 \\ u - \frac{a}{[\gamma(\gamma-1)]^{\frac{1}{2}}} & \frac{u^2}{2} & u + \frac{a}{[\gamma(\gamma-1)]^{\frac{1}{2}}} \end{bmatrix}, \quad (16)$$

$$\Lambda^a = [(\gamma - 1)/\gamma]^{\frac{1}{2}} \begin{bmatrix} -a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{bmatrix}. \quad (17)$$

We now split  $\Lambda^a$  into  $\Lambda^{a+}$  and  $\Lambda^{a-}$  given by

$$\Lambda^{a+} = [(\gamma - 1)/\gamma]^{\frac{1}{2}} \begin{bmatrix} 0 & 0 & 0 \\ 0 & \phi a & 0 \\ 0 & 0 & a \end{bmatrix}; \Lambda^{a-} = [(\gamma - 1)/\gamma]^{\frac{1}{2}} \begin{bmatrix} -a & 0 & 0 \\ 0 & -\phi a & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (18)$$

and then obtain the acoustic fluxes  $F^{a\pm}$  defined by

$$F^{a\pm} = R^a \Lambda^{a\pm} (R^a)^{-1}, \quad (19)$$

when mathematical manipulation in (17) are performed we obtain

$$F^{a\pm} = \begin{bmatrix} \pm[(\gamma - 1)/\gamma]^{\frac{1}{2}} \phi a & \rho \\ \frac{p}{2} \pm [(\gamma - 1)/\gamma]^{\frac{1}{2}} \phi a & \rho u \\ \frac{pu}{2} \pm \frac{pa}{2[(\gamma - 1)]^{\frac{1}{2}}} \pm [(\gamma - 1)/\gamma]^{\frac{1}{2}} \phi a & \frac{\rho u^2}{2} \end{bmatrix}. \quad (20)$$

Here  $\phi$  is a dissipation control parameter to be determined by requiring that the stability condition gives the best stability limit on the permissible time step. The variation of  $\phi$  with Mach number is given by the following expression (Balakrishnan & Deshpande 1994a):

$$\phi = \begin{cases} 1.37925 - 2.6774M - 4.34512M^2 & \text{for } M \leq 0.234355, \\ 0.520381 - 0.030885M & \text{for } 0.234355 < M \leq 10.373218, \\ 0.2 & \text{for } M > 10.373218. \end{cases} \quad (21)$$

The above expression for  $\phi$  are obtained by using a curve-fit for the stability plot. The linear stability analysis applied to the equation

$$\frac{\partial}{\partial t} + A^t \frac{\partial U}{\partial x} + A^{a+} \frac{\partial U}{\partial x} + A^{a-} \frac{\partial U}{\partial x} = 0, \quad (22)$$

is somewhat complex because the matrices  $A^t, A^{a\pm}$  do not commute and hence do not admit simultaneous diagonalisation. Following the von Neumann stability analysis which assumes that,

$$U^n(x) = \int \hat{U}(\xi) e^{2\pi i \xi x} d\xi. \quad (23)$$

We substitute for  $U^n(x)$  in the upwind scheme,

$$\frac{U_k^{n+1} - U_k^n}{\Delta t} + \frac{A_k^t (U_k^n - U_{k-1}^n)}{\Delta x} + \frac{A_k^{a+} (U_k^n - U_{k-1}^n)}{\Delta x} + \frac{A_k^{a-} (U_{k+1}^n - U_k^n)}{\Delta x} = 0, \quad (24)$$

for  $u > 0$  and,

$$\frac{U_k^{n+1} - U_k^n}{\Delta t} + \frac{A_k^t (U_{k+1}^n - U_k^n)}{\Delta x} + \frac{A_k^{a+} (U_k^n - U_{k-1}^n)}{\Delta x} + \frac{A_k^{a-} (U_{k+1}^n - U_k^n)}{\Delta x} = 0, \quad (25)$$

for  $u < 0$ . We obtain the amplification matrix  $G(\beta)$  in the form,

$$G = I - \frac{\Delta t}{\Delta x} Y; Y = E_T A^+ + E_B A^{a+} + E_F A^{a-}, \quad (26)$$

where  $E_T = E_B$  if  $u > 0$  and  $E_F$  if  $u < 0$ , and  $E_B$  and  $E_F$  are defined by

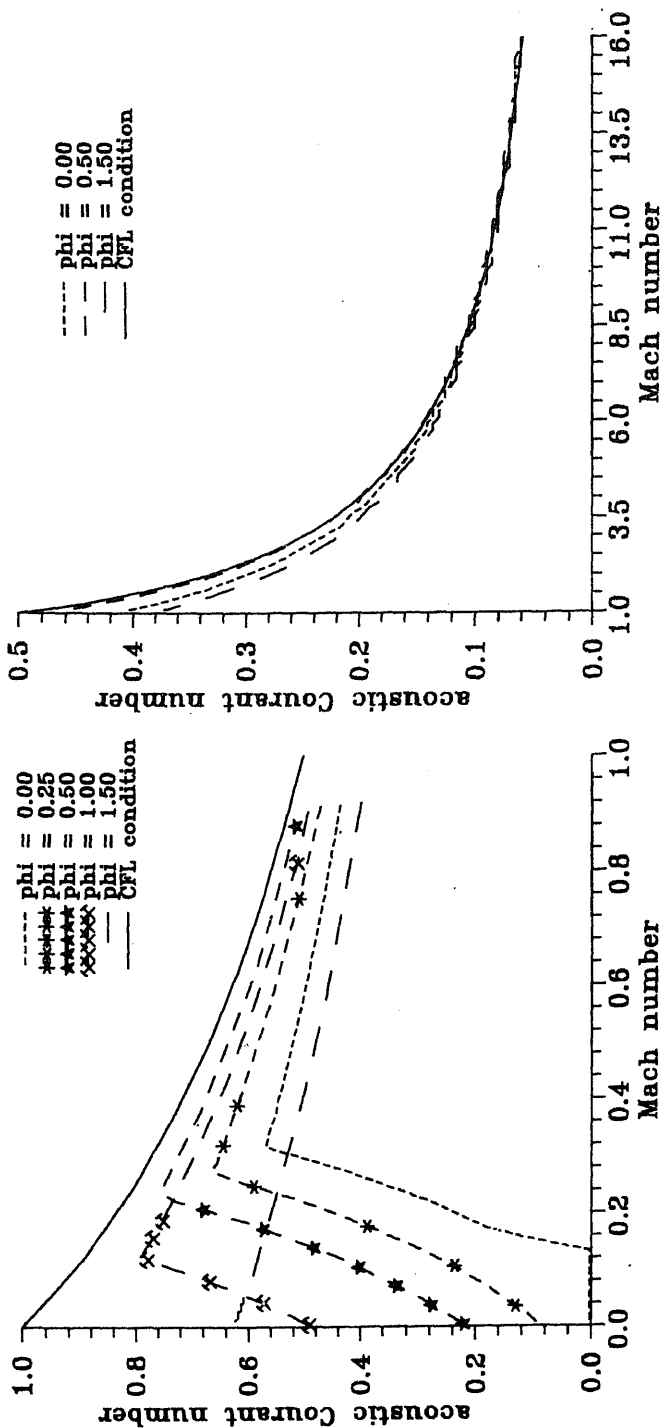
$$E_B = 1 - e^{-i\beta}, E_F = e^{i\beta} - 1, \beta = 2\pi\xi\Delta x. \quad (27)$$

The spectral theory of stability demands that the spectral radius of  $G$  be less than or equal to unity. Figure 1 shows the stability plots for the above scheme. These plots show the maximum allowed acoustic courant number ( $a(\Delta t/\Delta x)$ ) as a function of Mach number for values of dissipation control parameter  $\phi$  varying from 0 to 1.5. It is obvious that the Acoustic Flux Vector Splitting Scheme is unstable for very low Mach numbers when  $\phi = 0$ . We can increase the low mach number limit of the allowed  $a(\Delta t/\Delta x)$  by increasing  $\phi$ . Table 1 shows the real and imaginary parts of the eigenvalues of  $A^{\pm a}$ .

Table 1. Eigenvalues of  $A^{\pm a}$  for various values of  $\phi$

$\phi$	Eigenvalues of $A^{\pm a}$	
	Real parts	Imaginary parts
0.00	0.000000	0.000000
	$\pm 1.000000$	0.000000
	$\mp 0.250000$	0.000000
0.10	$\pm 0.046877$	0.000000
	$\pm 1.011320$	0.000000
	$\mp 0.158201$	0.000000
0.25	$\pm 0.090635$	0.000000
	$\pm 1.034360$	0.000000
	0.000000	0.000000
0.50	$\pm 0.202128$	$\pm 0.127213$
	$\pm 0.202128$	$\mp 0.127213$
	$\pm 1.095740$	0.000000
1.00	$\pm 0.455587$	$\pm 0.269326$
	$\pm 0.455587$	$\mp 0.269326$
	$\pm 1.338830$	0.000000
2.00	$\pm 0.797879$	$\pm 0.419213$
	$\pm 0.797879$	$\mp 0.419213$
	$\pm 2.154240$	0.000000

A factor of  $[(\gamma - 1)/\gamma]^{\frac{1}{2}}$  has been pulled out of the eigenvalues in the above table for convenience. It is obvious from this table that for  $\phi < 0.25$  all the eigenvalues of

Figure 1. Comparison of stability limits of AFVS scheme for different values of  $\phi$ .

$A^{a+}$  and  $A^{a-}$  are real but are of mixed sign. At  $\phi = 0.25$  all the eigenvalues of  $A^{a+}$  are positive and those of  $A^{a-}$  are negative. For values of  $\phi > 0.25$  the eigenvalues are complex. An interesting point that emerges from the present stability and eigenvalue analysis is that a stable upwind scheme based on split fluxes need not have all positive and all negative eigenvalues for the split flux Jacobians. This has been followed as an unnecessarily restrictive methodological principle in the past while designing FVS schemes.

## 2.2 Acoustic flux difference splitting (AFDS)

Acoustic flux difference splitting is yet another variant of the wave-particle splitting method. We start with the transport-acoustic split Euler equations (8),

$$\frac{\partial U}{\partial t} + \frac{\partial F^t}{\partial x} + \frac{\partial F^a}{\partial x} = 0 \quad (28)$$

as before but treat the term  $\partial F^a / \partial x$  differently (Balakrishnan and Deshpande 1991a, 1992). The transport term  $\partial F^t / \partial x$  is upwind differenced as before. The acoustic term  $\partial F^a / \partial x$  is discretised by using Flux Difference Splitting. There are many ways of achieving this discretisation. In the AFDS version studied here we use Roe-linearisation. We must hasten to add that Roe (1981) has treated the unsplit flux  $F$  in this manner while we use Roe's approach for the acoustic flux vector  $F^a$  only. For this purpose following Roe, we introduce the parameter vector,

$$w_1 = \sqrt{\rho} \quad ; \quad w_2 = \sqrt{\rho} u \quad ; \quad w_3 = p / \sqrt{\rho}, \quad (29)$$

which is somewhat different from the parameter vector used by Roe. In terms of the above parameter vector the conserved vector  $U$  and acoustic flux vector  $F^a$  become

$$U = \left[ \begin{array}{c} w_1^2 \\ w_1 w_2 \\ \frac{w_1 w_3}{(\gamma - 1)} + \frac{w_2^2}{2} \end{array} \right] \quad \text{and} \quad F^a = \left[ \begin{array}{c} 0 \\ w_1 w_2 \\ w_2 w_3 \end{array} \right]. \quad (30)$$

The corresponding averaged quantities are defined by,

$$\left. \begin{aligned} \bar{u} &= \frac{\bar{w}_2}{\bar{w}_1} = \frac{\sqrt{\rho_L} u_L + \sqrt{\rho_R} u_R}{\sqrt{\rho_L} + \sqrt{\rho_R}} \\ \bar{a}^2 &= \frac{\gamma \bar{w}_3}{\bar{w}_1} = \frac{\gamma (p_L / \sqrt{\rho_L} + p_R / \sqrt{\rho_R})}{\sqrt{\rho_L} + \sqrt{\rho_R}} \end{aligned} \right\}. \quad (31)$$

The averaged matrix  $\bar{A}^a$  is given by,

$$\bar{A}^a = \left[ \begin{array}{ccc} 0 & 0 & 0 \\ \frac{(\gamma - 1)}{2} \bar{u}^2 & -(\gamma - 1) \bar{u} & (\gamma - 1) \\ -\frac{\bar{a}^2 \bar{u}}{\gamma} + \frac{(\gamma - 1)}{2} \bar{u}^3 & \frac{\bar{a}^2}{\gamma} - (\gamma - 1) \bar{u}^2 & (\gamma - 1) \bar{u} \end{array} \right]. \quad (32)$$

It is interesting to observe that averaged  $\bar{A}^a$  matrix is very similar to  $A^a$  given by (11) except that  $u$  and  $a$  in (11) are replaced by  $\bar{u}$  and  $\bar{a}$  defined above. The matrix  $\bar{A}^a$  satisfies the well-known property,

$$\Delta F = \bar{A}^a \Delta U, \quad (33)$$

where  $\Delta F^a = F_R^a - F_L^a$ ,  $\Delta U = U_R - U_L$ . Because of the similarity between  $\bar{A}^a$  and  $A^a$  it follows that,

$$\bar{\lambda}(\bar{A}^a) = 0, \pm \bar{a}[(\gamma - 1)/\gamma]^{\frac{1}{2}}. \quad (34)$$

The corresponding right eigenvectors are,

$$\mathbf{r}_1 = \begin{bmatrix} 0 \\ 1 \\ \bar{u} - \frac{\bar{a}}{[\gamma(\gamma - 1)]^{\frac{1}{2}}} \end{bmatrix}; \mathbf{r}_2 = \begin{bmatrix} 1 \\ \bar{u} \\ \frac{\bar{u}^2}{2} \end{bmatrix}; \mathbf{r}_3 = \begin{bmatrix} 0 \\ 1 \\ \bar{u} + \frac{\bar{a}}{[\gamma(\gamma - 1)]^{\frac{1}{2}}} \end{bmatrix}. \quad (35)$$

The acoustic flux  $F_C^a$  at the cell interface are calculated using the formula

$$F_C^a = \frac{1}{2} \left[ (F_L^a + F_R^a) - \sum_{j=1}^3 \Delta \psi_j^a \bar{r}_j | \lambda_j | \right] \quad (36)$$

where  $L$  and  $R$  stand for the states to the left and right of the cell face  $C$  and  $\Delta \psi_j$  are given by,

$$\left. \begin{aligned} \Delta \psi_1^a &= \frac{\bar{\rho} \Delta u}{2} - \sqrt{\gamma/(\gamma - 1)} \frac{\Delta p}{2\bar{a}} \\ \Delta \psi_2^a &= \Delta \rho \\ \Delta \psi_3^a &= \frac{\bar{\rho} \Delta u}{2} + \sqrt{\gamma/(\gamma - 1)} \frac{\Delta p}{2\bar{a}} \end{aligned} \right\} \quad (37)$$

and,

$$\Delta \rho = \rho_R - \rho_L, \Delta u = u_R - u_L, \Delta p = p_R - p_L. \quad (38)$$

The total flux at a cell face  $C$  is then given by the sum of upwind interpolated transport flux  $F^t$  and  $F_C^a$ . As a last comment on AFDS we may add that even though AFDS is based on our wave-particle idea it is somewhat different from AFVS described before. Both AFDS and AFVS treat the transport part of the flux in the same fashion but acoustic part  $F^a$  on a cell face is computed differently by these methods.

### 3. Peculiar velocity based upwind (PVU) method

Here we discuss yet another line of approach for construction of novel algorithms for Euler equations of gas dynamics. This line of attack called the PVU method, is an application of the moment method strategy so successfully applied by Deshpande (1986c) and Mandal (1989) and Mandal & Deshpande (1994) to develop the Kinetic Flux Vector Splitting (KFVS) method. The KFVS method, which has been recently surveyed by Deshpande (1993), has turned out to be extremely robust in numerically solving a variety of 2-D and 3-D problems arising in inviscid gas dynamics. The KFVS method suffers from three basic deficiencies :

- (i) like many other flux vector splitting methods it is highly diffusive. This property is both a virtue and a vice. The KFVS method owes its robustness to its highly diffusive character (virtue) and *ipso facto* leads to unacceptably large smearing of contact discontinuity and boundary layers (vice) ;
- (ii) the KFVS method assumes a rest frame because the splitting is accomplished by dividing the molecular velocity into positive half ( $v > 0$ ) and negative half ( $v < 0$ ) ;
- (iii) the integration of the Maxwellian distribution over  $v > 0$  and  $v < 0$  in KFVS

leads to formulae involving error functions whose computation for every mesh point can be expensive.

Raghurama Rao & Deshpande (1991a, 1995) have recently advanced the concept of splitting based on the peculiar velocity  $c$  (also called the thermal velocity in the jargon of the Kinetic Theory of Gases). The split flux expressions so obtained are free of the defects (ii) and (iii) above, and lead to the PVU method which is more efficient than the KFVS method while at the same time possessing its robust property.

We will now describe the basis of the PVU method (Raghurama Rao & Deshpande (1991a, 1995)). Consider the 1-D Boltzmann equation

$$\frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = J(f, f), \quad (39)$$

where  $v$  is the molecular velocity,  $f$  is the velocity distribution function,  $J(f, f)$  is the collision term whose structure is of no concern here as it vanishes in the Euler limit. The basic unknown  $f$  in (35) is a function of time  $t$ , position  $x$  and velocity  $v$ . The Maxwellian velocity distribution denoted by  $F$  is given by

$$F = \frac{\rho}{I_0} (\beta/\pi)^{\frac{1}{2}} \exp \left[ -\beta (v - u)^2 - \frac{I}{I_0} \right], \quad (40)$$

where  $\rho$  is the mass density,  $\beta = \frac{1}{2RT}$ ,  $T$  is the temperature,  $R$  is the gas constant per unit mass,  $u$  is the fluid velocity,  $I$  is the internal energy variable corresponding to nontranslational degrees of freedom (this variable is required to force a given value of  $\gamma$  for the gas),  $I_0$  is the average internal energy due to nontranslational degrees of freedom, given by  $I_0 = [(3 - 2\gamma)/(2(\gamma - 1))] RT$  and  $\gamma$  is the ratio of specific heats. One of important properties of

$$\frac{\partial F}{\partial t} + v \frac{\partial F}{\partial x} = 0, \quad (41)$$

is that the Euler equations of motion,

$$\frac{\partial U}{\partial t} + \frac{\partial G}{\partial x} = 0, \quad (42)$$

can be cast in the compact form

$$\langle \Psi, \frac{\partial F}{\partial t} + v \frac{\partial F}{\partial x} \rangle = 0, \quad (43)$$

where

$$\Psi = \text{moment function vector} = \begin{bmatrix} 1 \\ v \\ I + \frac{v^2}{2} \end{bmatrix}, \quad (44)$$

$$U = \langle \Psi, F \rangle = \int_0^\infty dI \int_{-\infty}^\infty dv \Psi F = \begin{bmatrix} \rho \\ \rho u \\ \rho E \end{bmatrix}, \quad (45)$$



$$G = \langle \Psi, F \rangle = \int_0^\infty dI \int_{-\infty}^\infty dv \Psi v F = \begin{bmatrix} \rho u \\ p + \rho u^2 \\ p u + \rho u E \end{bmatrix}. \quad (46)$$

The above connection between the Boltzmann equation (37) and the Euler equations (38) is at the root of many kinetic schemes. Raghurama Rao & Deshpande (1991a, 1995) rewrite (37) in the form ,

$$\frac{\partial F}{\partial t} + \frac{\partial}{\partial x} (uF) + \frac{\partial}{\partial x} (cF) = 0, \quad (47)$$

where  $c = v - u$  is the peculiar velocity. Taking  $\Psi$ -moments of (43) gives

$$\frac{\partial U}{\partial t} + \frac{\partial G^t}{\partial x} + \frac{\partial G^a}{\partial x} = 0, \quad (48)$$

where

$$G^t = \langle \Psi, uF \rangle = \int_0^\infty dI \int_{-\infty}^\infty dc \Psi u F, \quad (49)$$

$$G^a = \langle \Psi, cF \rangle = \int_0^\infty dI \int_{-\infty}^\infty dc \Psi c F. \quad (50)$$

The flux vectors  $G^t$  and  $G^a$  defined by (44) and (45) are exactly the same as  $F^t$  and  $F^a$  given by (9). It is interesting to observe that the physical arguments leading to the equation (8) involving splitting of the flux vector into the transport and the acoustic parts, are different from those behind equation (44). Even then both ways of looking at the splitting of the flux vector lead to identical expressions for the transport and the acoustic parts. The basic idea behind writing the Boltzmann equation (39) in the form given by (45) is the recognition that the  $u$ -part is unidirectional while the  $c$ -part is multidirectional. Another way of looking at this difference is that  $u$  is a deterministic variable while the random variable  $c \sim N(0, 1/(2\beta))$ , that is,  $c$  follows a Normal distribution with zero mean and variance equal to  $1/(2\beta)$ . The motion of a particle can be thought of as consisting of an orderly motion ( $u$ ) and a random motion ( $c$ ) due to thermal agitation of molecules. The PVU method recognises this difference in the behaviour of  $u$ -term and  $c$ -term and treats them differently. It is also believed that this way of dealing with the two terms is useful in constructing genuinely multidimensional upwind schemes.

The next question in the development of the PVU method is the upwind differencing of  $\partial G^t/\partial x$  and  $\partial G^a/\partial x$  terms. So far as the transport term is concerned we follow the same method as before, that is, we write

$$\frac{\partial}{\partial x} (uF) = \frac{\partial}{\partial x} \left( \frac{u + |u|}{2} F \right) + \frac{\partial}{\partial x} \left( \frac{u - |u|}{2} F \right), \quad (51)$$

and then obtain the upwind differenced approximation as

$$\begin{aligned} \left[ \frac{\partial}{\partial x} (uF) \right]_j &= \frac{1}{\Delta x} \left[ \left( \frac{u + |u|}{2} F \right)_j - \left( \frac{u + |u|}{2} F \right)_{j-1} \right] \\ &+ \frac{1}{\Delta x} \left[ \left( \frac{u - |u|}{2} F \right)_{j+1} - \left( \frac{u - |u|}{2} F \right)_j \right]. \end{aligned} \quad (52)$$

After taking  $\Psi$ -moments as per the moment method strategy (Deshpande 1986c) we obtain the upwind differencing at the Euler level as

$$\left(\frac{\partial G^i}{\partial x}\right) = \left[\frac{\partial}{\partial x}(uF)\right]_j = \frac{1}{\Delta x} \left[ \left(\frac{u+|u|}{2}F\right)_j - \left(\frac{u+|u|}{2}F\right)_{j-1} \right] + \frac{1}{\Delta x} \left[ \left(\frac{u-|u|}{2}F\right)_{j+1} - \left(\frac{u-|u|}{2}F\right)_j \right]. \quad (53)$$

Let us now deal with the upwind differencing of the acoustic term  $\partial G^a/\partial x$ . Following Raghurama Rao & Deshpande (1991a, 1995) we observe that

$$\begin{aligned} \frac{\partial G^a}{\partial x} &= \langle \Psi, \frac{\partial}{\partial x}(cF) \rangle \\ &= \langle \Psi, \frac{\partial}{\partial x} \left( \frac{c+|c|}{2}F \right) \rangle + \langle \Psi, \frac{\partial}{\partial x} \left( \frac{c-|c|}{2}F \right) \rangle \\ &= \frac{\partial}{\partial x} \langle \Psi, \frac{c+|c|}{2}F \rangle + \frac{\partial}{\partial x} \langle \Psi, \frac{c-|c|}{2}F \rangle \end{aligned} \quad (54)$$

We therefore obtain split acoustic fluxes  $G^{a+}$  and  $G^{a-}$  given by

$$G^{a+} = \langle \Psi, \frac{c+|c|}{2}F \rangle \text{ and } G^{a-} = \langle \Psi, \frac{c-|c|}{2}F \rangle. \quad (55)$$

This splitting is based on the peculiar velocity  $c$  and hence this scheme is called Peculiar Velocity based Upwind (PVU) scheme. In terms of the split fluxes (48) gives

$$\frac{\partial G^a}{\partial x} = \frac{\partial G^{a+}}{\partial x} + \frac{\partial G^{a-}}{\partial x}. \quad (56)$$

Performing the integration with respect to  $c$  and  $I$  in the formulae (49) we obtain

$$G^{a\pm} = \left[ \begin{array}{c} \pm \frac{\rho}{2\sqrt{\pi\beta}} \\ \frac{p}{2} \pm \frac{\rho u}{2\sqrt{\pi\beta}} \\ \frac{pu}{2} \pm \frac{1}{2\sqrt{\pi\beta}} \left( \frac{p}{2} + \rho E \right) \end{array} \right]. \quad (57)$$

Both the AFVS and PVU schemes are somewhat similar to the CUSP (Convection Upwind and Split Pressure) scheme of Jameson (1993) in the sense that the convection terms in the Euler equations are upwind differenced and the pressure terms are split. However, the CUSP scheme is different from the present methods. The transport flux and the acoustic flux of AFVS and PVU methods are not the same as the convective flux and pressure flux either in the CUSP scheme of Jameson or that due to Denton (1983). Hence, the eigenvalues of the corresponding flux Jacobians are also different for CUSP and the present methods. The method of upwind differencing the convective and pressure fluxes is also different in the AFVS and PVU methods compared to the CUSP method. Let us now compare the expressions for the split acoustic fluxes for the AFVS and PVU methods. We note that  $\beta = \gamma/(2a^2)$

and hence  $G^{a\pm}$  for the PVU method reduce to

$$G^{a\pm} = \begin{bmatrix} \pm \frac{\rho a}{\sqrt{2\pi\gamma}} \\ \frac{p}{2} \pm \frac{\rho u a}{\sqrt{2\pi\gamma}} \\ \frac{pu}{2} \pm \frac{\rho a}{\sqrt{2\pi\gamma}} \left( \frac{p}{\rho} \frac{\gamma+1}{2(\gamma-1)} + \frac{u^2}{2} \right) \end{bmatrix}. \quad (58)$$

Now compare these with the AFVS expression (18)

$$G^{a\pm} = \begin{bmatrix} \pm \sqrt{\frac{\gamma-1}{\gamma}} \phi \rho a \\ \frac{p}{2} \pm \sqrt{\frac{\gamma-1}{\gamma}} \phi \rho u a \\ \frac{pu}{2} \pm \rho a \sqrt{\frac{\gamma-1}{\gamma}} \left( \frac{p}{\rho(\gamma-1)} + \frac{\phi u^2}{2} \right) \end{bmatrix}. \quad (59)$$

We observe that PVU formulae for  $G^{a\pm}$  given by (52) are very similar to those of AFVS given by (18). However, they are *not* identical even if we try to choose a suitable value of the dissipation control parameter  $\phi$  in (18). For, a comparison between (18) and (52) suggests that

$$\phi = \frac{1}{[2\pi(\gamma-1)]^{\frac{1}{2}}}. \quad (60)$$

This value when substituted in expressions (18) yields

$$G^{a\pm} = \begin{bmatrix} \pm \frac{\rho a}{\sqrt{2\pi\gamma}} \\ \frac{p}{2} \pm \frac{\rho u a}{\sqrt{2\pi\gamma}} \\ \frac{pu}{2} \pm \frac{\rho a}{\sqrt{2\pi\gamma}} \left( \frac{p}{\rho} \frac{2\pi}{\gamma-1} + \frac{u^2}{2} \right) \end{bmatrix}, \quad (61)$$

which shows that the split fluxes for mass and momentum are in agreement with the respective expressions for the PVU method but the split fluxes for energy do not match. We therefore conclude that the AFVS and PV-splitting are allied ideas but they do not lead to identical expressions for the split fluxes even when the parameter  $\phi$  is adjusted. It appears that one more parameter may be required for obtaining identical formulae. Whether this is possible within the framework of AFVS remains to be seen.

Finally we end this section by stating that the PV-splitting can be easily extended to multidimensions. We start with the 2-D Boltzmann equation

$$\frac{\partial F}{\partial t} + \frac{\partial}{\partial x} (v_1 F) + \frac{\partial}{\partial x} (v_2 F) = 0. \quad (62)$$

The collision term gets dropped as it becomes zero in the Euler limit. The 2-D Maxwellian  $F$  is given by

$$F = \frac{\rho}{I_0} \frac{\beta}{\pi} \exp \left[ -\beta (v_1 - u_1)^2 - \beta (v_2 - u_2)^2 - \frac{I}{I_0} \right], \quad (63)$$

$$I_0 = \frac{2-\gamma}{\gamma-1} RT. \quad (64)$$

Following the same procedure as before we rewrite (62) in the form

$$\frac{\partial F}{\partial t} + \frac{\partial}{\partial x} (u_1 F) + \frac{\partial}{\partial x} (u_2 F) + \frac{\partial}{\partial x} (c_1 F) + \frac{\partial}{\partial x} (c_2 F) = 0, \quad (65)$$

and define the transport and acoustic fluxes by

$$G_1^t = \langle \Psi, u_1 F \rangle, \quad G_2^t = \langle \Psi, u_2 F \rangle, \quad G_1^a = \langle \Psi, c_1 F \rangle, \quad G_2^a = \langle \Psi, c_2 F \rangle, \quad (66)$$

The split acoustic fluxes are then given by

$$G_1^{a\pm} = \langle \Psi, \frac{c_1 \pm |c_1|}{2} F \rangle, \quad G_2^{a\pm} = \langle \Psi, \frac{c_2 \pm |c_2|}{2} F \rangle. \quad (67)$$

Performing the integrations in the formulae (60) with respect to  $I$ ,  $c_1$  and  $c_2$  we obtain

$$G_1^{a\pm} = \left[ \begin{array}{c} \pm \frac{\rho}{2\sqrt{\pi\beta}} \\ \frac{p}{2} \pm \frac{\rho u_1}{2\sqrt{\pi\beta}} \\ \pm \frac{\rho u_2}{2\sqrt{\pi\beta}} \\ \frac{pu_1}{2} \pm \frac{\rho}{2\sqrt{\pi\beta}} \left( \frac{p}{\rho} \frac{\gamma+1}{2(\gamma-1)} + \frac{u_1^2 + u_2^2}{2} \right) \end{array} \right], \quad (68)$$

$$G_2^{a\pm} = \left[ \begin{array}{c} \pm \frac{\rho}{2\sqrt{\pi\beta}} \\ \pm \frac{\rho u_1}{2\sqrt{\pi\beta}} \\ \frac{p}{2} \pm \frac{\rho u_2}{2\sqrt{\pi\beta}} \\ \frac{pu_2}{2} \pm \frac{\rho}{2\sqrt{\pi\beta}} \left( \frac{p}{\rho} \frac{\gamma+1}{2(\gamma-1)} + \frac{u_1^2 + u_2^2}{2} \right) \end{array} \right]. \quad (69)$$

The above formulae are based on splitting along  $x$ - $y$  directions of a Cartesian frame. When PVU is used in the finite volume framework, the fluxes must be determined on a cell face not necessarily parallel to either of the coordinate directions. On an arbitrary cell face the split fluxes are given by

$$G^{a\pm} = \left[ \begin{array}{c} \pm \frac{\rho}{2\sqrt{\pi\beta}} \\ n_1 \frac{p}{2} \pm \frac{\rho u_1}{2\sqrt{\pi\beta}} \\ n_2 \frac{p}{2} \pm \frac{\rho u_2}{2\sqrt{\pi\beta}} \\ \frac{pu_n}{2} \pm \frac{\rho}{2\sqrt{\pi\beta}} \left( \frac{p}{2\rho} + E \right) \end{array} \right], \quad (70)$$

where  $n_1$  and  $n_2$  are the direction cosines of the outward normal on the cell face,  $u_n$  is the fluid velocity normal to the cell face and  $E$  is the internal energy per unit mass. We observe that the split flux formulae  $G^{a\pm}$  of the PVU method do not

involve error functions and exponential terms and hence are expected to lead to a more efficient computational method than the KFVS method. Also, no rest frame is assumed in deriving the above formulae. The next section discusses the results obtained by solving a large number of 2-D problems using AFVS, AFDS and PVU methods.

## 4. Results and discussions

For some preliminary results obtained using the AFDS method reference is made to Balakrishnan & Deshpande (1991). AFDS method has been found to be less dissipative and less robust than the AFVS method. Here we concentrate on the results obtained by using AFVS and PVU methods.

Both the AFVS and PVU schemes have been tested on a large number of 2-D problems (Raghurama Rao(1994) and Balakrishnan and Deshpande (1992). It is a standard practice in Computational Fluid Dynamics to evaluate the performance of new schemes by computing subsonic, transonic and supersonic flows for GAMM/AGARD test cases. We give below a few examples demonstrating the capability of these two new methods. More details are available in Raghurama Rao (1994) and Balakrishnan & Deshpande (1992). The AFVS method was applied to compute flow past NACA0012 airfoil on  $128 \times 64$  O - type structured grid with outer boundary five chord lengths away from the mid-chord point. The free stream Mach number is 0.85 and the angle of attack is  $1^\circ$ . This is an AGARD test case. The pressure and Mach contours obtained are shown in figure 2. The contours are smooth and are indicative of the ability of the scheme to capture shocks accurately.

The PVU method was also applied to the above test example. The computations were done on an unstructured mesh with adaptation. High resolution finite volume version of PVU method was used (Raghurama Rao 1994). The starting mesh, the adapted mesh as well as the pressure contours obtained on these meshes are shown in figures 3, 4, 5 and 6. Only a part of the mesh is shown for clarity. The outer boundary is located 15 chord lengths away from the airfoil, which is not shown in the above figures. The shocks on the upper and the lower surfaces are very accurately captured. Table 2 shows the lift and drag coefficients ( $C_L$  and  $C_D$ ) obtained by AFVS and PVU schemes. Also shown in this table are the standard values of these coefficients in the AGARD report (AGARD 1986).

**Table 2.** Lift and drag coefficients obtained with AFVF and PVU schemes.

Case $M_\infty \alpha$	AGARD range		AFVS		PVU	
	$C_L$	$C_D$	$C_L$	$C_D$	$C_L$	$C_D$
0.85 $1^\circ$	0.330-0.3889	0.0464-0.0590	0.4145	0.0607	0.3351	0.0565

The coefficients predicted by more accurate computations using PVU method fall within the AGARD range. The AFVS computations were not done with mesh adaptation and the coefficients predicted by the AFVS method are slightly outside the AGARD range. More accurate computations done by using AFVS method push these values into the AGARD range.

At the outer limit of the speed range is the hypersonic flow involving blunt body shock. The performance of the AFVS scheme was further tested by computing

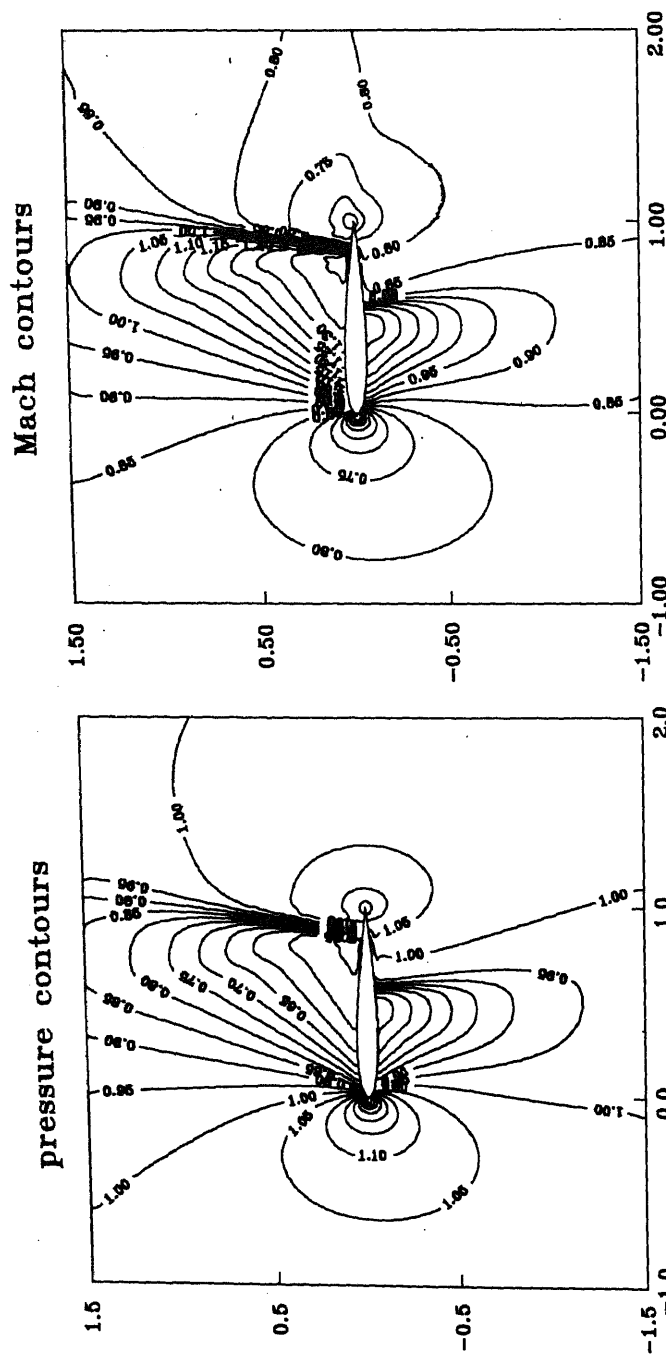
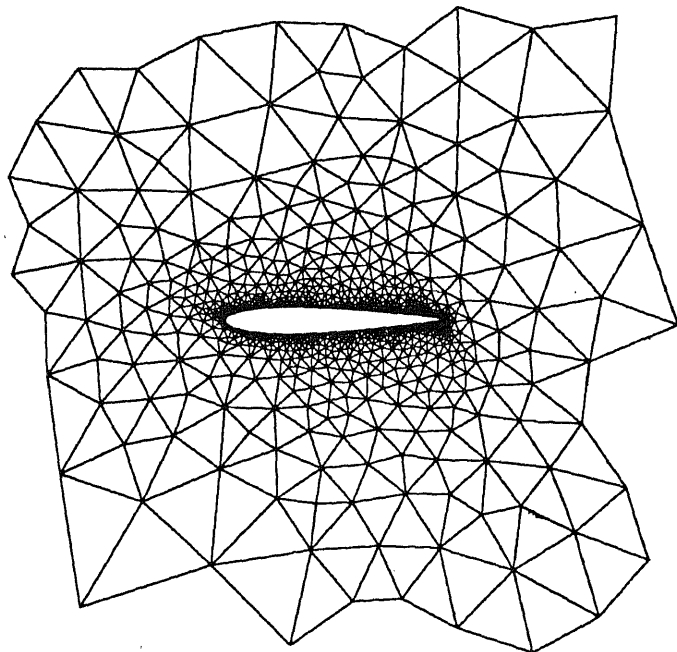
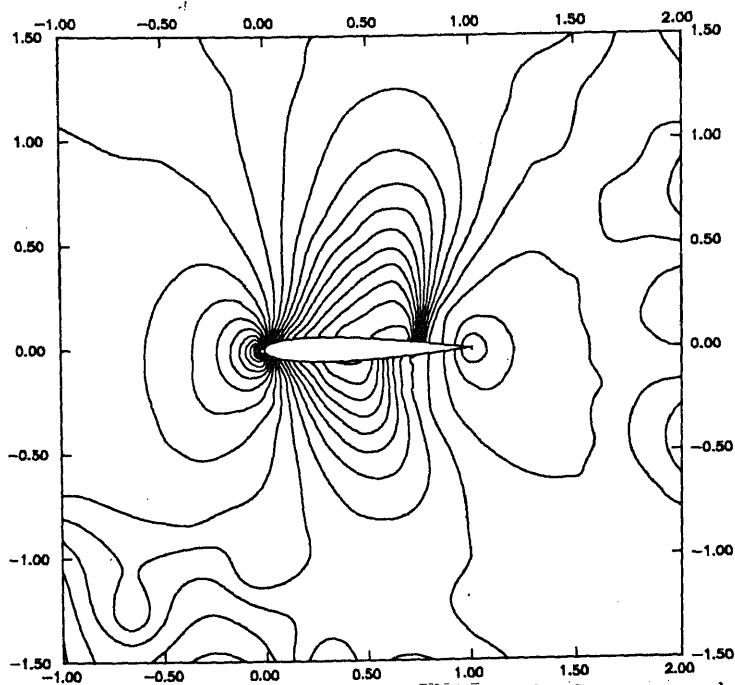


Figure 2. Pressure and Mach contours for AFVS scheme (AGARD Test Case, NACA 0012, 128 x 64 grid, free stream Mach no. = 0.85,  $\alpha = 1.0^\circ$ ).



**Figure 3.** Starting mesh for NACA0012 airfoil, points = 1123, cells = 2056, edges = 3179.



**Figure 4.** Pressure contours obtained on the starting mesh. Contour levels from 0.54 to 1.5 (increment 0.04), scheme: High resolution PVU finite volume method, Mach no. = 0.85, angle of attack =  $1^\circ$ .

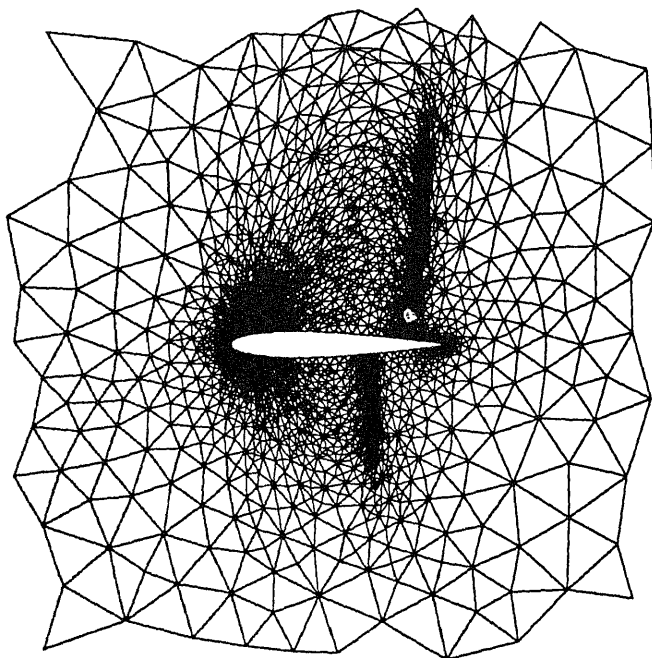


Figure 5. Adapted mesh for NACA0012 airfoil, points = 5724, cells = 11197, edges = 16921.

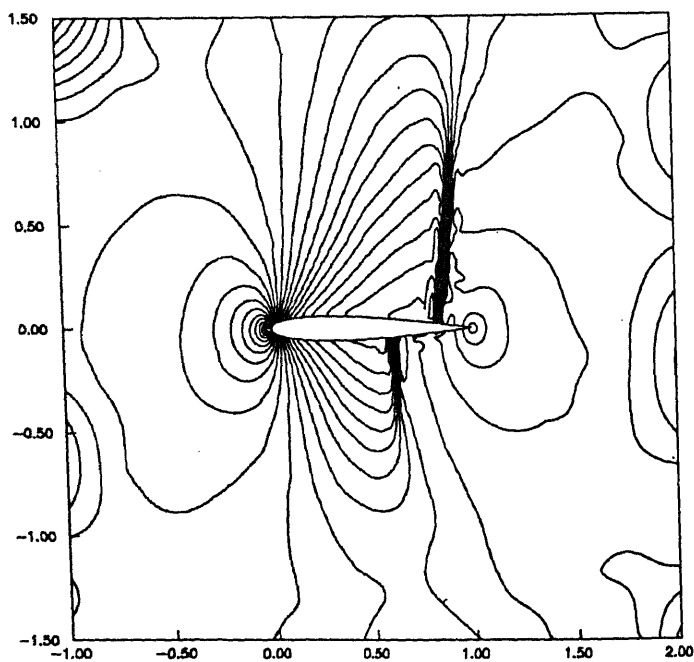


Figure 6. Pressure contours obtained on the adapted mesh. Contour levels from 0.48 to 1.5 (increment 0.04), scheme: High resolution PVU finite volume method, Mach no. = 0.85, angle of attack =  $1^\circ$ .



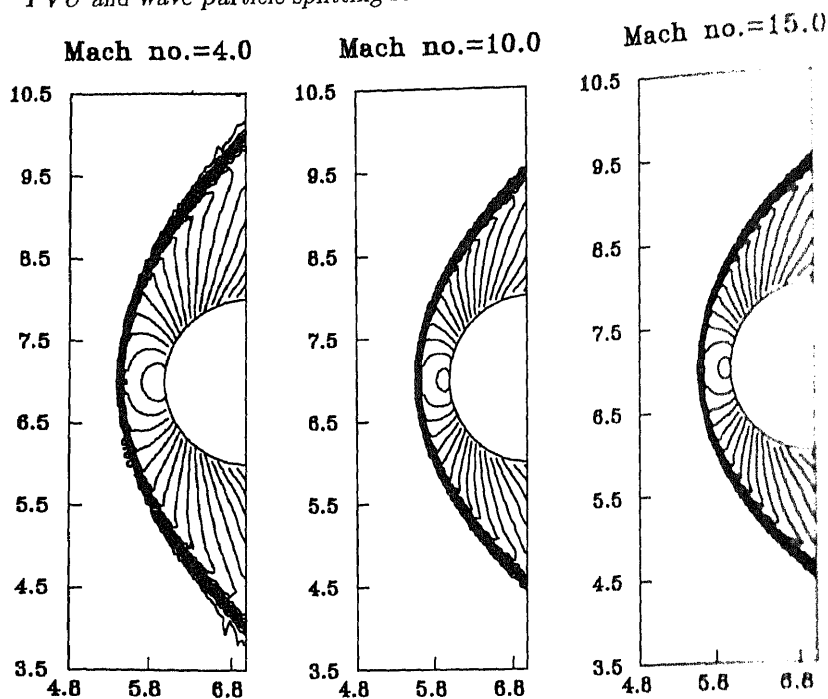


Figure 7. Pressure contours for AFVS scheme (flow past cylinder,  $99 \times 98$  grid, MUSCL, contours 0.05 units apart).

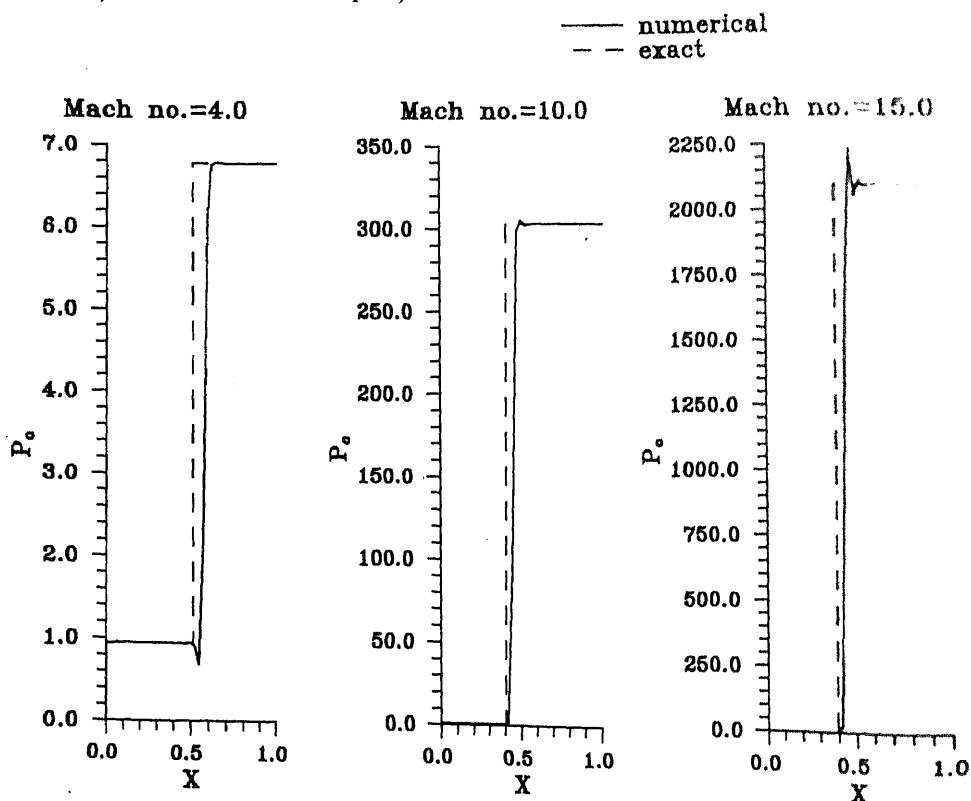


Figure 8.  $P_0$  variation along the stagnation line for AFVS scheme (flow past a cylinder,  $99 \times 98$  grid, MUSCL).

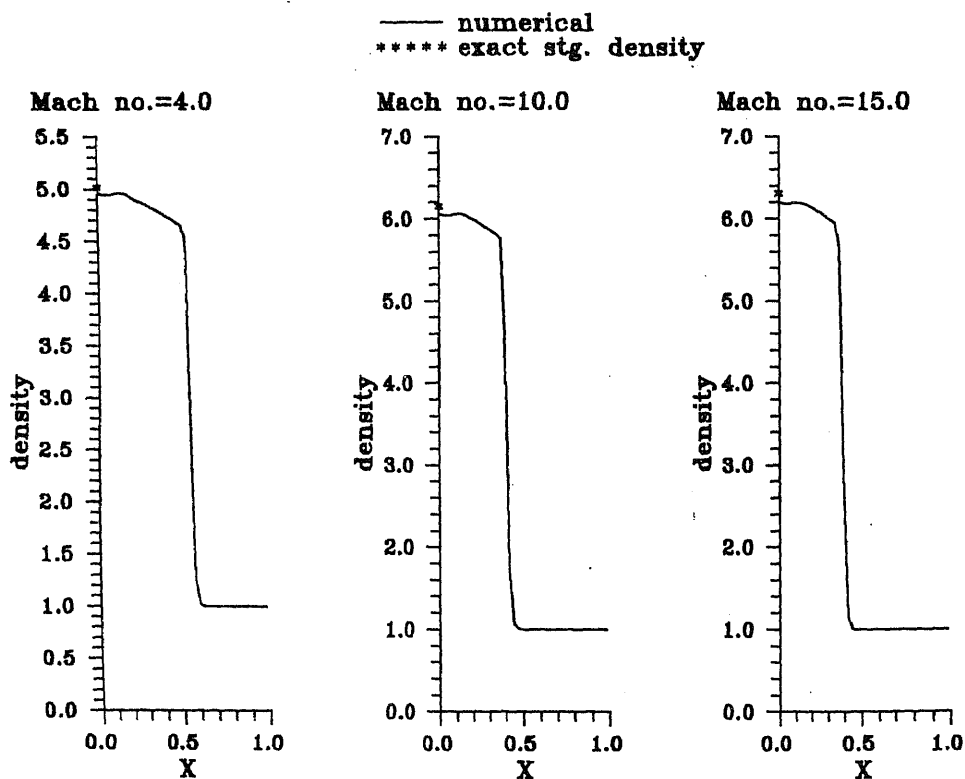


Figure 9. Density variation along the stagnation line for AFVS scheme (flow past a cylinder, 99 x 98 grid, MUSCL).

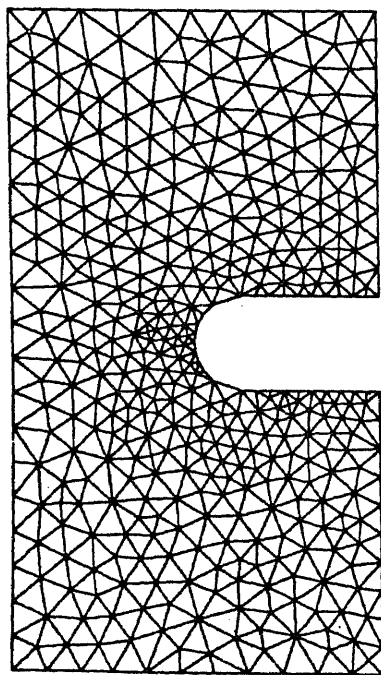
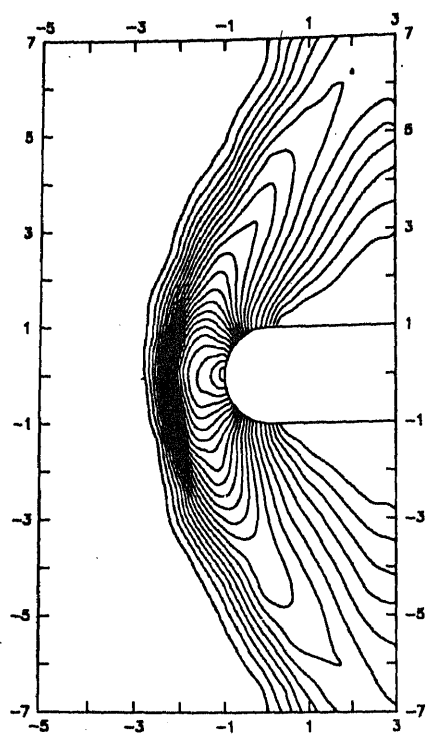
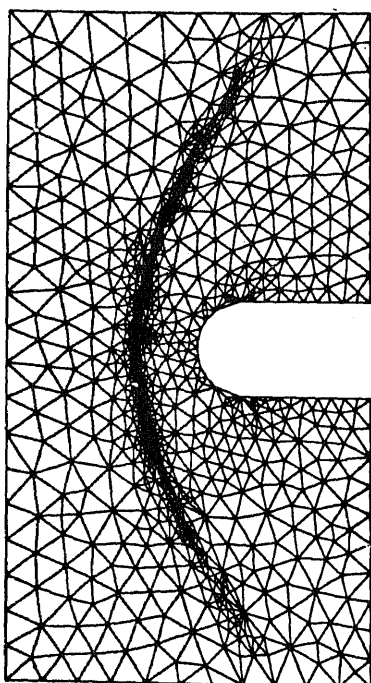


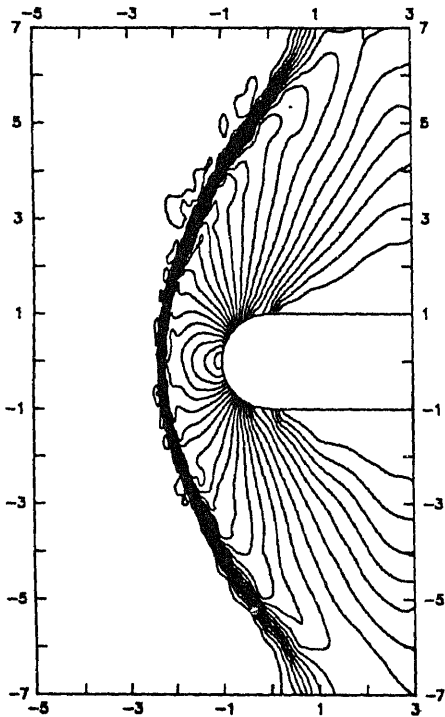
Figure 10. Starting mesh for blunt body, points = 446, cells = 800, edges = 1245.



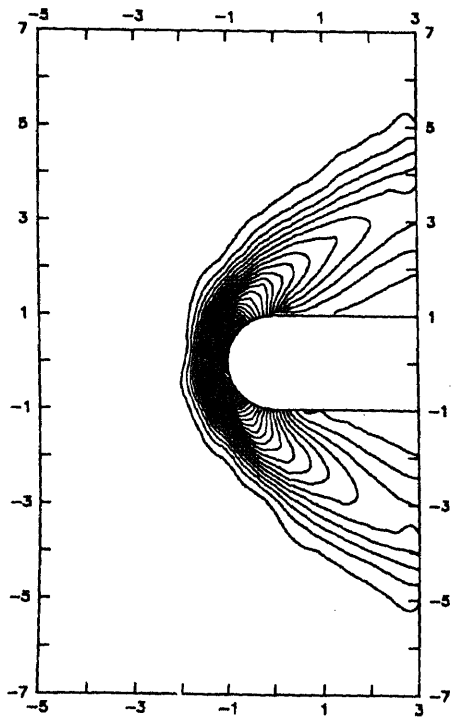
**Figure 11.** Pressure contours on starting mesh. Contour levels from 1.2 to 5.5 (increment 0.2), Mach no. = 2.0, angle of attack =  $0^\circ$ , scheme: High resolution PVU finite volume method.



**Figure 12.** Adapted mesh for supersonic flow, points = 1017, cells = 1942, edges = 2958.



**Figure 13.** Pressure contours on adapted mesh. Contour levels from 0.7 to 5.7 (increment 0.2), Mach no. = 2.0, angle of attack =  $0^\circ$ , scheme: High resolution PVU finite volume method.



**Figure 14.** Pressure contours on starting mesh. Contour levels from 2.0 to 78.0 (increment 2.0), Mach no. = 8.0, angle of attack =  $0^\circ$ , scheme: First order PVU finite volume method.

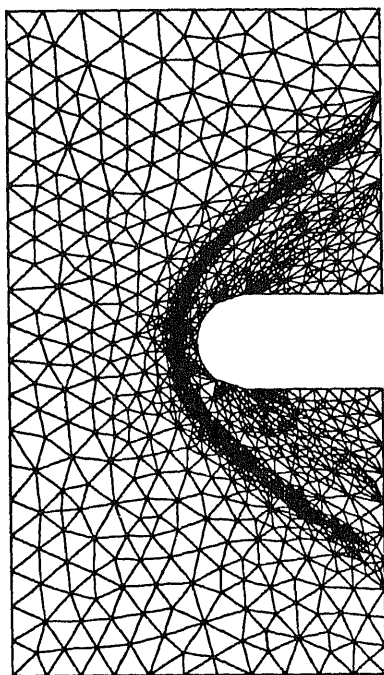


Figure 15. Adapted mesh for hypersonic flow, points = 1520, cells = 2940, edges = 4459.

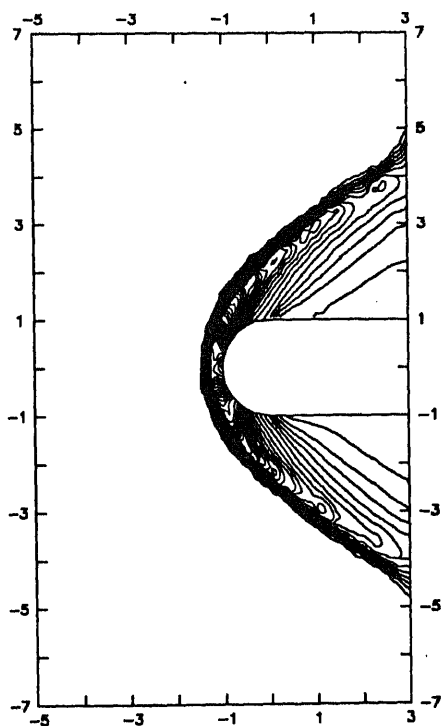


Figure 16. Pressure contours on adapted mesh. Contour levels from 2.0 to 82.0 (increment 2.0), Mach no. = 8.0, angle of attack =  $0^\circ$ , scheme: First order PVU finite volume method.

flow past a semi-cylinder. The computations were made for three different different mach numbers : 4, 10 and 15. A uniform  $99 \times 98$  structured grid is employed for computations. The radius of the outer boundary is chosen to be 3.5 times the radius of the semi-cylinder. The pressure contours given by the AFVS scheme for three different Mach numbers, the total pressure variation and the density variation along the stagnation line are shown in figures 7,8 and 9. The smoothness of these contours demonstrate the robustness of the AFVS scheme in capturing strong shocks.

The PVU scheme was applied to the test problem given by Arminjon & Dervieux (1993) for hypersonic flow. The computations are done for  $M = 2.0$  and  $M = 8.0$  with zero angle of attack using high resolution and first order PVU scheme respectively on an unstructured mesh. Again the starting mesh, adapted mesh and the pressure contours obtained on these meshes are shown in figures 10 to 16. The bow shock is closer to the body in hypersonic case compared to the supersonic flow. This is a well-known result of gas dynamics. The shocks are captured crisply by the PVU method with grid adaptation.

Extensive computations performed using the AFVS and PVU schemes show the basic soundness of the wave-particle idea and splitting based on the peculiar velocity of molecules. The basic idea of having a discrete mathematical model inheriting as many physical properties of the fluid flow as possible appears quite sound and turns out to be often promising needing further study. Implicit in the AFVS method is the physically meaningful model that fluid behaves partly particle like and partly wave like. The PVU method is based on a different but closely related idea that the motion of particles is a random motion superimposed on a unidirectional motion. The particle like behaviour in wave-particle splitting is equivalent to the unidirectional motion in the PVU method. The random motion taking place in all directions is similar to the wave spreading in all directions. It would be very tempting to design a new scheme taking the multidirectionality of the pressure part into account. The genuinely multidimensional upwind scheme based on the Boltzmann equation has been developed by Raghurama Rao & Deshpande (1991b) and Eppard & Grossman (1993). The results show a lot of improvement but these schemes are at present quite expensive. These two schemes also do not separate unidirectional and random motion of the molecules as in the PVU scheme. Further work is required to exploit these ideas in a more efficient manner.

## References

- AGARD 1986 Test cases for inviscid flow field methods. Report of Fluid Dynamics Panel Working Group 07, AGARD-AR 211
- Arminjon P, Dervieux A 1993 Construction of TVD-like artificial viscosities in two-dimensional arbitrary FEM grids. *J. Comput. Phys.* 106: 176-198
- Balakrishnan N, Deshpande S M 1991 New upwind schemes with wave-particle splitting for inviscid compressible flows. 91 FM 11, Fluid Mechanics Report, Dept. Aerosp. Eng., Indian Institute of Science, Bangalore
- Balakrishnan N, Deshpande S M 1992a Assessment of the performance of wave-particle splitting with other flux vector and flux difference splitting schemes. 92 FM 7, Fluid Mechanics Report, Dept. Aerosp. Eng., Indian Institute of

Science, Bangalore (Also presented in IMACS symp. held in Bangalore, India, December-1992)

- Balakrishnan N, Deshpande S M 1994a New upwinding scheme based on the wave-particle splitting *Computational fluid dynamics*, '94. *Proc. Second European CFD Conference*. (eds) S Wagner, E H Hirschel, J Periaux, R Piva (Chichester: John Wiley & Sons) pp 1-8
- Balakrishnan N, Deshpande S M 1995 New upwind method exploiting the wave-particle behaviour of fluid flow. *Comput. Fluid Dyn. J.* (in press)
- Balakrishnan N, Raghurama Rao S V 1992b New flux vector and flux difference splitting methods for Euler equations of gas dynamics. *Proc. 5th Asian Cong. on Fluid Mechanics* Taejon, Korea
- Denton J D 1983 An improved time-marching method for turbomachinery flow calculations. *J. Eng. Gas Turbines Power* 105
- Deshpande S M 1986a On the Maxwellian distribution, symmetric form and entropy conservation for Euler equations. NASA-TP-2583
- Deshpande S M 1986b Kinetic theory based new upwind methods for inviscid compressible flows. AIAA Paper 86-0275
- Deshpande S M 1993 Boltzmann schemes for continuum gas dynamics. *Sadhana* 18: 405-430
- Eppard W M, Grossman B 1993 A Multidimensional kinetic-based upwind solver for Euler equations. AIAA Paper No. 93-3303 CP
- Jameson A 1993 Artificial diffusion, upwind biasing, limiters and their effect on accuracy and multigrid convergence in transonic and hypersonic flows. AIAA Paper 93-3359
- Jameson A, Schmidt W, Turkel E 1981 Numerical solution of Euler equations finite volume method using Runge Kutta time stepping. AIAA-81-1259
- Liou M S, Steffen J Jr 1991 A new flux-splitting scheme. NASA TM 104404, NASA Lewis Research Centre, Cleveland, Ohio
- MacCormack R W 1990 Solution of the Navier-Stokes equation in three dimensions. AIAA Paper No. 90-1520
- Mandal J C 1989 *Kinetic upwind method for inviscid compressible flows*. Ph D thesis, Dept. Aerosp. Eng., Indian Institute of Science, Bangalore
- Mandal J C, Deshpande S M 1994 Kinetic flux vector splitting for Euler equations. *Comput. Fluids* 23: 447-478
- Osher S, Solomon F 1982 Upwind difference schemes for hyperbolic system of conservation laws. *Math. Comput.* 38: 339
- Quirk J J 1992 A contribution to the great Reimann solver debate. ICASE Report NO. 92-64

- Raghurama Rao S V 1994 New upwind methods based on kinetic theory for inviscid compressible flows. Ph D thesis, Dept. Mech. Eng., Indian Institute of Science, Bangalore
- Raghurama Rao S V, Deshpande S M 1991a A class of efficient kinetic upwind methods for compressible flows. 91 FM 12, Fluid Mechanics report, Dept. Aerosp. Eng., Indian Institute of Science, Bangalore
- Raghurama Rao S V, Deshpande S M 1991b A genuinely multidimensional upwind Boltzmann scheme for Euler equations. 91 FM 6, Fluid Mechanics Report, Dept. Aerosp. Eng., Indian Institute of Science, Bangalore
- Raghurama Rao S V, Deshpande S M 1994 Peculiar velocity based upwind method on unstructured meshes. *Computational fluid dynamics '94. Proc. Second European CFD Conference.* (eds) S Wagner, E H Hirschel, J Periaux, R Piva (Chichester: John Wiley & Sons) pp. 89-96
- Raghurama Rao S V, Deshpande S M 1995 Peculiar velocity based upwind method for inviscid compressible flows. *Comput. Fluid Dyn. J., Jpn.* (to appear)
- Roe P L 1981 Approximate Riemann solvers, parameter vectors and difference schemes. *J. Comput. Phys.* 43: 357-372
- Sanders R H, Prendergast K H 1974 The possible relation of the 3-kilospace arm to the explosions in the galactic nucleus. *Astrophys. J.* 188: 489-500
- Steger J L, Warming R F 1981 Flux vector splitting of inviscid gas dynamic equations with application to finite difference methods. *J. Comput. Phys.* 40: 263-293
- van Leer B 1982 Flux vector splitting for Euler equations. *Lecture Notes in Physics* 170: 507
- van Leer B 1990 Flux vector splitting for 1990's. Invited Lecture for the CFD Symposium on Aeropropulsion, NASA Lewis Research Center, NACA CP-3078



# Sonochemical reaction engineering

K S GANDHI and R KUMAR

Department of Chemical Engineering, Indian Institute of Science, Bangalore 560 012, India

**Abstract.** Ultrasound has been widely used by chemists to enhance yields as well as rates of homogeneous as well as heterogeneous chemical reactions. The effect of ultrasound on the course of chemical reactions is mediated through cavitation bubbles it generates. High temperatures and pressures are attained inside the cavitating bubbles when they collapse. The extreme conditions so generated lead to the formation of reactive intermediates, e.g., free radicals, inside the bubbles, which cause chemical reactions to occur when they enter the surrounding liquid. This is the mechanism through which ultrasound influences the path of homogeneous reactions. The cavitation bubbles collapse asymmetrically in the vicinity of solids, e.g., catalyst particles. Asymmetric collapse lead to formation of high speed microjets. The microjets can enhance transport rates, the increase surface area through pitting as well as particle fragmentation through collisions. Both can alter the rates of heterogeneous reaction rates. It however appears that these effects do not exhaust the scope of the influence of ultrasound on heterogeneous reactions. Modelling and quantitative prediction of the effect of ultrasound on chemical reactions is however at a stage of infancy as the phenomena are complex. Only a few examples of modelling exist in literature. Apart from this, reactor design and scaleup pose significant problems. Thus sonochemical reaction engineering offers large scope for research and development efforts.

**Keywords.** Sonochemistry; ultrasound; cavitation; sonolysis; sonoreactors.

## 1. Introduction

Sounds audible to the human ear are of frequencies between 16 Hz and 16 kHz. Sound of frequency beyond 16 kHz is called ultrasound. In this range also there are two regions viz. high frequency and power ultrasound. High frequency ultrasound, of between 1 and 10 MHz, finds extensive use in medical diagnosis and chemical analysis. It is only the power ultrasound however that can induce chemical reactions or enhance their rates.

Power ultrasound, normally considered to fall in the frequency range of 20 to 100

kHz, already finds extensive use in the cleansing of components, emulsification of immiscible liquids, plastic welding and machining of hard, brittle materials etc. The special characteristic of this range of frequencies is that its amplitude can be made reasonably large, resulting in high intensity of sound. It is this high intensity sound encountered in the power ultrasound range, which shows interesting, and many a time unusual, chemical effects.

The use of power ultrasound, referred to as ultrasound or sound from here on, in laboratory scale syntheses of chemicals has shown a significant spurt during the last two decades. The results of such practices have been very impressive in many cases. Thus, application of ultrasound has been found to (i) yield entirely different products from the same reactants, (ii) enhance the rates of the reactions, and (iii) permit, at very mild reaction conditions, syntheses of products that can normally be obtained only under very extreme conditions of pressure and temperature. Such chemical effects have been reported not only for homogeneous reactions, but also where liquid-liquid and liquid-solid interfaces are involved. This brings in a new technique and a new kind of reactor, the sonochemical reactor, to the domain of chemical engineering. In this paper, we present a review of the kinds of reactions which can be influenced by power ultrasound, the mechanisms of reaction rate enhancement, the meager attempts at modelling made upto this time and the kind of larger reactors which are likely to be eventually employed.

## 2. Origin of effect of ultrasound

Unlike light, sound waves do not interact directly with molecules to cause chemical reactions, because their wavelengths are very high. For example, when ultrasound of frequency 20 to 10,000 kHz passes through water at a velocity of  $1500 \text{ m s}^{-1}$ , the wavelengths fall between 7.5 and 0.015 cm. These are too large to have any direct implication on sonochemical effects. Further, the energy contained in sound waves is too small compared to the chemical bond energies. With a large number of reactions having been studied now, the origin of chemical reactions has been recognized as arising out of cavitation.

Pure liquids possess very high tensile strengths and can withstand very high negative pressures before breaking down and generating vapour bubbles. If a bubble of this kind has to be generated in pure water, a negative pressure of the order of 1000 atmospheres is required. Unless special precautions are taken, liquids always contain small amounts of dissolved gases. These pockets of gas constitute weak spots in the liquid. When a sound wave is passed through a medium, it generates alternate cycles of compression and rarefaction. If during rarefaction, the pressure at one of these weak spots falls below the vapour pressure of the medium, bubbles be get generated. Thus, generation of vapour-filled gas bubbles, or cavitation, can occur at very small negative pressures, in most cases less than one atmosphere. The presence of weak spots or nuclei in the liquid thus facilitates bubble formation. The nuclei can also be formed out of very tiny gas microbubbles present in the liquid, arising out of collapsing cavities, a phenomenon that will be discussed a little later. Yet another source of nuclei are the gas cavities present on any tiny particles of solid present in the medium.

The behaviour of these cavities in varying pressure fields is of great interest to engineers and has been studied widely. One of the important conclusions is that, in

response to oscillating pressure fields, some nuclei can grow relatively slowly to fairly large sizes and collapse in the subsequent few microseconds. As the growth phase is very slow, it is reasonable to assume that enough heat exchange between the bubble contents and the surrounding liquid takes place to ensure isothermal conditions. The growth phase occurs during the rarefaction cycle of the pressure wave. Due to inertia of the liquid, the bubble grows for some more time during the compression cycle also. However, as the compression portion of the cycle proceeds, the bubble surface reverses its direction of motion and its radius starts decreasing as a function of time. This reduction in radius caused by the compressive pressure field is also aided by the inertia of the inward rushing liquid. Thus the decrease in the radius can be extremely fast. As the time of collapse is very short, the amount of heat transferred between the bubble and the surrounding liquid is negligible and the collapse can be treated as adiabatic. The last phases of the bubble collapse therefore correspond to adiabatic compression. Under these conditions, the temperatures and pressures attained in the bubble at its collapse can be high. For example, a temperature of the order of 1500 K and 100 10 MN/m<sup>2</sup> can be easily achieved. Recently, Flint & Suslick (1991), by analysing the sonoluminescence spectra obtained by sonicating silicone oil have reported that the temperature attained by cavitation bubbles for the system under investigation was  $5075 \pm 156$  K. Thus cavitating bubbles can create pockets of high temperature and pressures in an otherwise cool liquid.

The very final stages of a collapsing bubble, when the velocity of its surface exceeds the velocity of sound, are not charted out clearly. It is possible that such a cavity fragments into small bits, which can serve as nuclei for further cavitation subsequently.

Another interesting phenomenon is also associated with bubble collapse. When a bubble collapses near a solid wall or near another bubble, the motion of the bubble would not be spherically symmetric. Asymmetric bubble collapse is quite predominant near the solid boundary. Such collapse gives rise to tiny jets of liquid, about a hundred microns in diameter and having a velocity of between 100 and 500 m/s. The lifetime of these 'microjets' is of the order of a few microseconds. These microjets can cause intense convection, and enhance the mass transfer rates tremendously.

Apart from cavity growth and collapse, there are other phenomena associated with cavitation. When the collapse occurs over a few cycles, it can happen that during each cycle, the bubble wall velocity may cross the velocity of sound. Thus when the bubble starts to expand again, its velocity is higher than the velocity of sound in the liquid medium. This gives rise to shock waves. The shock waves can partially clean the surface of the solid, and also increase the mass transfer rates. Further, the fracture as well as pitting of larger particles is the direct result of microjets and shock waves. When the bubble collapses near the surface, it yields a microjet which impinges on the surface causing either erosion or fracture. When the cavity rebounds, it produces a shock wave which adds to this erosion. In both cases the elastic limit of the solid is crossed and erosion occurs in the plastic regime. Lush *et al* (1983) have given an excellent analysis, where both the phenomena are taken into account. Photographs of the severe damage caused to marine propellers by these causes are prominently displayed in texts on classical fluid mechanics.

The result of fracture and pitting is obviously the increase of surface area available for reaction. But pitting and fracture can also activate the surface by physically

dislodging any passivating layer.

Shock waves and pressure fluctuations can lead to emulsification in liquid-liquid systems leading to increase in interfacial areas.

Thus all the chemical and physical effects of ultrasound have to be explained in terms of the effects of cavitating bubbles. Before embarking on such a task, it is worthwhile to review the wide variety of influence of ultrasound.

### 3. Nature of changes brought about by power ultrasonication

Power ultrasound has been found to bring about remarkable changes in both physical and chemical processes. Some of the categories of changes brought about by sonication are described below.

#### 3.1 Physical processes

A number of physical processes which are significantly influenced by ultrasonication have been described in literature. Some of these are described below.

- (a) *Degassing*: Ultrasound can quickly degas solvents and this technique has been used extensively. It is likely that rectified diffusion plays a role in this effect.
- (b) *Emulsification*: Two immiscible liquids are easily emulsified by subjecting them to high power ultrasonication. This method is already industrially employed. Apart from the tensile stresses exerted by the pressure field which can cause drop breakup, shock waves emitted by collapsing bubble also contribute to fragmentation of drops and hence to emulsification.
- (c) *Crystallization*: Both crystallization and precipitation get significantly modified by power ultrasound. Martin (1993) refers to work done at Harwell which has shown that ultrasound applied at the nucleation stage in a crystallizer yields much cleaner, higher quality crystals. Carrying out crystallization in a continuous fluidised bed with ultrasound applied at the base yields more uniform-sized crystals. It is likely that the convection caused by the microjets plays a role in these phenomena.
- (d) *Cell disruption*: High intensity ultrasound is routinely used in biological laboratories for cell disruption. These effects are perhaps akin to the emulsifying effect of ultrasound that has already been described.
- (e) *Surface cleaning*: A large number of investigators have reported that when particles of metals and other materials are subjected to ultra sonication, they may break and their surface gets pitted. It is very likely that microjets created by collapsing bubbles have a strong role here. This instant cleaning of the surface and generation of new surface area has profound influence on overall rates of reaction.
- (f) *Particle fusion*: It has been reported by Suslick (1989) that under high intensity power ultrasonication, particles suspended in the medium move towards each other at very high velocities (of the order of 150 m/s). When these particles collide with each other, so much heat is generated that local melting occurs and particles get welded to each other. Once again, microjets seem to be the cause of the high velocities attained by the particles.

- (g) *Agglomeration*: Use of ultrasound leading to coagulation of small particles has been reported, and can aid in prevention of pollution. The convection caused by ultrasound is likely to increase the collisions between particles and may lead to increased rates of coagulation.

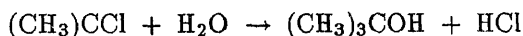
Though we have postulated qualitative explanations for the various effects mentioned above, quantitative modelling of these phenomena is quite intricate and, hardly any reports of such efforts exist in literature.

Apart from the physical effects, ultrasonication brings about interesting and useful chemical effects. These effects are so numerous and all-pervading, that a new branch of chemistry called '*sonochemistry*' has emerged during the last two decades.

### 3.2 Chemical effects

That ultrasound can bring about reaction where one least expects it is brought out by the sonolysis of water itself. If water is ultrasonicated, say in a cleaning bath, small quantities of hydrogen peroxide can be detected. This is easily seen if a KI solution in water is sonicated. The KI yields  $I_2$  which can be easily analysed by standard methods. This reaction has been studied in detail (Hart & Henglein 1985), though it has no industrial relevance. Some other examples of changes brought out by sonication are mentioned below.

- (a) *Acceleration of reaction rates*: The major portion of applications of sonochemistry deals with acceleration of reaction rates. Normally, these would involve reactions with two phase viz. liquid/liquid or liquid/solid. For example, hydrogenation of alkenes by nickel powder is enormously accelerated ( $> 10^5$ fold) by ultrasonication. Recently, even a dehydrogenation reaction (e.g., of tetrahydronaphthalene on palladium catalyst) has been found to be accelerated by application of ultrasonication. Perhaps these effects can be attributed to the convection due to microjets created by ultrasonication in the case of hydrogenation and degassing in the case of dehydrogenation. The literature is replete with other examples of acceleration of heterogeneous reactions when sonication is employed. Another typical example is the reduction of an  $\alpha, \beta$ -unsaturated ketone to the corresponding allylic alcohol through the use of zinc and acetic acid. This reaction is extremely slow with stirring and cannot be taken to completion even after prolonged reaction time. However, with ultrasonication, the reaction gives quantitative yield of the product in one hour at room temperature. Once again one might surmise that microjets play a role in this situation as well. However, there are situations where a homogeneous reaction can also be accelerated which is harder to rationalize. A typical example is the hydrolysis of tertiary butyl chloride, in 50% aqueous ethanol to give tertiary butyl alcohol.

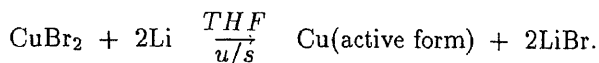


Against stirring, this reaction shows an increase in rate by a factor of 20 through sonication (at  $10^\circ C$ ).

- (b) *Avoidance of purification of reagents and solvent*: Grignard reagents are extremely useful in organic syntheses and involve the reaction of an alkyl or aryl

halide with magnesium in diethyl ether (as solvent). Normally, this reaction is conducted with stirring, and it is absolutely essential to use purified magnesium and distilled dry ether. Similar is the case when the reagent is used to carry out the desired reaction. Without these precautions, the formation of Grignard reagent, or its reaction with substrates, cannot be accomplished. With sonication, it has been found possible to use Mg without any prior treatment and commercial sample of ether. Similarly, the non-nucleophilic strong base lithium diisopropyl amide could be prepared starting with lithium metal itself under the influence of ultrasonication, while such a preparation was not possible otherwise. Thus, sonication permits us to avoid, in some circumstances, a number of purification steps, which in an industry can be economically beneficial. While it is easy to speculate that degassing caused by ultrasonication may be playing a role here, it is hard to contend that the role of ultrasonication is confined to only this.

- (c) *Replacing the role of phase transfer catalysts*: Phase transfer catalysts are popular among organic chemists for conducting two-phase reactions, when the reactants are insoluble in each other. The phase transfer catalyst carries the reagent from one phase to another, where it reacts easily. A number of reactions have been reported where the same reactions can be performed in the absence of phase transfer catalysts, but in the presence of ultrasonication. One such example is the cyclopropanation of styrene with NaOH and  $\text{CHCl}_3$  (Repie & Vogt 1982). While conversions of only 30% were reached under stirring for 16 hours, 96% conversions could be reached with ultrasound in only one hour. Here it is hard to even speculate upon the role of ultrasonication.
- (d) *Use of less hazardous reagents, and milder conditions*: Active metal powders, required as catalysts, are normally prepared by reducing metal halides with potassium metal under reflux conditions and using a solvent like tetra hydro furan (THF). As potassium is difficult to handle, the process is hazardous particularly for large scale operation. When attempts are made to substitute K with less hazardous metal like Li, the reaction does not proceed. However, sonication forces the reaction to proceed with Li in less than forty minutes at room temperature (Boudjouk *et al* 1986),



Once again it is difficult to guess the role being played by ultrasonication, except to implicate the removal of oxide films from the lithium surface by the pitting action of ultrasonication.

- (e) *Change of reaction products (ultrasonic switching)*: Perhaps the most surprising influence, from the view point of rationalization, is that for some reactions, the products obtained by ultrasonication are entirely different from the ones obtained by simple stirring. For example, when benzyl bromide is reacted with toluene containing solid KCN and solid  $\text{Al}_2\text{O}_3$ , with stirring at higher temperature ( $\sim 130^\circ\text{C}$ ), we obtain a product of Friedel Crafts reaction viz., a benzyl substituted toluene. However, if the same reactants are sonicated

at room temperature, the reaction takes an entirely different course giving a product of substitution of bromide by cyanide viz. benzyl-cyanide.

Similarly, when a primary alcohol is treated with nitric acid, we obtain different products under stirring and ultrasonication. If the reactants are only stirred, the product obtained is an alkyl nitrate. If, however, the same reactants are sonicated, we obtain a product of oxidation leading to the formation of the corresponding carboxylic acid.

These effects are indeed very difficult to explain till some competing reactions between the reactants, one of which is accelerated by ultrasound, can be discerned.

- (f) *Preparation of amorphous metal powders:* Amorphous metal alloys called metal glasses find use in magnetic storage, power transformer cores etc. Further, some of them can act as catalysts. Suslick *et al* (1991) have reported the preparation of amorphous iron powder by ultrasonication of iron pentacarbonyl (20 kHz frequency,  $100 \text{ W cm}^{-2}$  intensity). The iron powder thus produced is highly porous in nature. The authors find that the powder is ten-fold more active than the commercial powder for the hydrogenation of carbon monoxide (the Fischer – Tropsch process) to low molecular weight alkanes.
- (g) *Formation and degradation of polymers:* Price *et al* (1992) report polymerization of methyl methacrylate under high intensity ultrasound and mention conditions under which polymers of controlled molecular weight, polydispersity and tacticity could be prepared. The polymerization reaction is perhaps initiated by free radicals which makes this reaction similar to the sonolysis of water mentioned earlier. They also report the degradation of PMMA (in methyl butyrate) through ultrasound. Here it is possible that the elongational stresses generated by the microjets lead to degradation.

Apart from the above mentioned examples, ultrasound is known to have a strong influence on electrochemical reactions (Mason *et al* 1990) and at lower intensities has shown useful effects in various aspects of biotechnology (Sinisterra 1992)

Thus, it can be seen that ultrasound can have a variety of effects on chemical reactions, and in many cases it is hard to even qualitatively explain the effects. Yet, the effects are often beneficial that an attempt to quantify these is worth the effort.

#### 4. Hypotheses about the mechanism of effect of ultrasound on homogeneous reactions

We consider at first homogeneous reactions. By this we mean those reactions where only a single phase is present, ignoring the cavitation bubbles. Though there is general agreement that sonochemical effects have their origin in cavitation, the actual fashion in which cavitation brings about the effects has been interpreted differently. There are two main ideas which have been used to qualitatively explain the reactions due to sonication. One of them considers that electrical discharge occurs during cavity collapse whereas the other one considers the reactions as occurring as a direct result of adiabatic collapse of the bubble generating small zones of high temperatures and pressures. In this paper, we shall confine ourselves to the latter only.

All the chemical effects of sonolysis can be directly attributed to these high temperatures and pressures. One of the views is that a thin liquid shell around the cavity also attains high temperatures and the reaction proceeds there as well. Ley & Low (1989) quote an example of the collapse of a cavitating bubble in heptane/decane where the vapour cavity attained a temperature of 5200 K and the thin shell of liquid around it attained a temperature of 1900 K. The thickness of the shell is however calculated to be only a few nanometres. It is however quite likely that the heat transfer during collapse may be controlled by the gas phase and the liquid surface may be remaining at essentially ambient conditions.

The fact that high temperatures are generated in a collapsing cavity can qualitatively explain some of the homogeneous reactions. For example, the formation of  $\text{H}_2\text{O}_2$  has been explained through this mechanism. Under the conditions of extremely high temperatures and pressures obtaining in the collapsing cavity, the  $\text{H}_2\text{O}$  molecules will dissociate giving rise to OH radicals which, when released into the liquid phase, combine together to form  $\text{H}_2\text{O}_2$ .

## 5. Modelling of homogeneous sonochemical reactors

A model has recently been attempted to describe a homogeneous reaction i.e., formation of  $\text{I}_2$  from KI solution in a batch sonochemical reactor (Prasad Naidu *et al* 1994). When a solution of KI is sonicated in a cleaning bath or through an ultrasonic horn, measurable quantities of  $\text{I}_2$  are liberated. A number of investigators have qualitatively studied this reaction, but the most comprehensive study has been conducted by Hart & Henglein (1985). They irradiated aqueous solutions of KI in a batch reactor with 300 kHz ultrasound under argon, oxygen and Ar -  $\text{O}_2$  mixtures of varying composition. They found the products formed to be  $\text{I}_2$  and  $\text{H}_2\text{O}_2$ . They also conducted experiments in the presence of ammonium molybdate which catalyses the oxidation of iodide ion by  $\text{H}_2\text{O}_2$  to iodine. Their main findings with regard to formation of  $\text{I}_2$  from KI are:

- (a) The rate of liberation of iodine increases with KI concentrations, but for a given KI concentration it remains constant with time.
- (b) Different rates of iodine formation are obtained when different gasses like  $\text{N}_2$ ,  $\text{O}_2$  and Ar are used.
- (c) When mixtures of Ar and  $\text{O}_2$  are used as dissolved gasses in the KI solution, the reaction rate passes through a maximum at an intermediate composition.

The model of Prasad Naidu *et al* (1994) tries to explain such results quantitatively on the following general framework.

### 5.1 Physical description of the model

The basic model is made up of three components viz. the formation, growth and collapse of bubbles, the reactions taking place during collapse of bubbles and the reactions taking place in the liquid which has been assumed to be well mixed. The KI solution is assumed to have a large number of nuclei in the form of small bubbles or gas entrapped in the crevices of the reactor wall. When the liquid medium is subjected to ultrasonication, these nuclei grow and collapse. Most of them grow



and collapse as transient cavities if the frequencies employed are lower than 100 kHz (Arakeri & Chakraborty 1990), as was the case in their work. These cavities grow and collapse in around one acoustic cycle. The actual phenomena of both heat and mass transfer during both the growth and collapse of a bubble is horrendously complex. Hence the authors have made use of a simple analysis proposed by Flynn (1964). In this analysis it is assumed that the entire growth phase and a part of the collapse phase can be treated as isothermal whereas the rest of the collapse phase can be treated as adiabatic. The transition from isothermal to adiabatic during the collapse phase is taken to occur when the internal gas pressure becomes equal to the vapour pressure of the liquid at the bulk temperature. The adiabatic collapse phase is assumed to end when the bubble wall velocity reaches the sonic velocity in the liquid medium. During the adiabatic collapse phase, the temperature and pressure inside the bubble increase due to compression, yielding extreme conditions. This can result in the formation of free radicals from the water vapour and oxygen, (if it is present in the dissolved gas). Here, the authors assume that at the end of the collapse phase the contents of the bubble attain the equilibrium composition dictated by thermodynamics. The composition of the gas bubble at the end of the collapse phase can be calculated without considering the kinetics of the various reactions taking place in the gas phase. The various radicals released into the medium undergo several reactions, one of them being the oxidation of iodide ion. How these radicals mix with the liquid phase and react is uncertain. The reactions are so fast that diffusion may be controlling. However because of the microjets and shock waves, the mixing can be significant. Further, on collapse, the bubble may fragment and yield a number of bubbles. In the absence of a clear picture, the model assumes complete mixing in the vessel. Thus, the various components of the model involve the following.

- (a) Growth and collapse of cavity,
- (b) evaluation of temperature and pressure at the end of the collapse,
- (c) calculation of equilibrium compositions in the gas phase at the end of collapse phase, and
- (d) material balances for various species in the liquid phase in which the gas contents are assumed to mix instantaneously.

The expansion and collapse phase are governed by the bubble dynamics equation, normally referred to as the Rayleigh-Plesset equation and is given by Plesset (1949).

$$R\ddot{R} + (3/2)\dot{R}^2 = (1/P_L)[P_L - P_\infty]. \quad (1)$$

The liquid pressure at the bubble surface is related to the inner pressure by

$$P_L(R) = P_i - 2\sigma/R. \quad (2)$$

Here  $P_i$  is the internal pressure of the bubble and is equal to  $P_g + P_v$ , the sum of the partial pressures of gas and vapour respectively. Though the gas partial pressure changes inside the bubble as its radius changes, its amount inside the bubble remains the same. Hence

$$P_g = P_{g0}(R_0/R)^{3\alpha}. \quad (3)$$

The external pressure is time varying and is characteristic of the applied acoustic field. Thus

$$P_{\infty} = P_b - P_a \sin \omega t. \quad (4)$$

Substituting (2), (3) and (4) in (1) and rearranging, we obtain

$$\frac{d\dot{R}}{dt} = \frac{1}{P_L} \left[ P_{g0} \left( \frac{R_o^{3\alpha}}{R^{3\alpha+1}} \right) + \frac{P_v}{R} - \frac{2\sigma}{R^2} - \frac{(P_b - P_a \sin \omega t)}{R} \right] - \frac{3\dot{R}^2}{2R}, \quad (5)$$

where  $\dot{R}$  represents  $dR/dt$ .

Equation (5) indicates that the bubble dynamics equation depends on  $\alpha$ , the value of which differs for isothermal and adiabatic phases. Thus, for the isothermal phase, we solve (5) with the following initial conditions and by setting  $\alpha = 1$  and  $P_v = P_S$ .

$$\text{At } t = 0, \quad R = R_o \quad \text{and} \quad \dot{R} = 0. \quad (6)$$

The end of this phase also means the beginning of the collapse phase. Following Flynn (1964), the collapse phase begins when  $P_g$  becomes equal to  $P_S$ . The value of bubble radius  $R_2$  at the beginning of collapse phase is readily calculated from  $P_S$  by

$$P_g = P_S = P_{g0} (R_o/R_2)^3. \quad (7)$$

During the second stage, no heat or mass transfer is permitted between the bubble and the surroundings. Therefore the pressure and volume of the bubble during this phase were related using the adiabatic gas laws. The bubble dynamics equation for  $R \leq R_2$  becomes:

$$\frac{d\dot{R}}{dt} = \frac{1}{\rho_L} \left[ 2P_S \left( \frac{R_2^{3\alpha}}{R^{3\alpha+1}} \right) - \frac{2\sigma}{R^2} - \frac{(P_b - P_a \sin \omega t)}{R} \right] - \frac{3\dot{R}^2}{2R}. \quad (8)$$

The bubble collapse is assumed to end when the bubble wall velocity reaches the velocity of sound in the liquid medium. Its size then is denoted by  $R_f$ . The initial gas pressure  $P_{g0}$  is related to initial cavity size by.

$$P_S + P_{g0} = P_b + 2\sigma/R_o. \quad (9)$$

The temperature and the pressure at the end of collapse are calculated from  $R_2$  and  $R_f$  as:

$$T_f = T_2 (R_2/R_f)^{3(\gamma-1)}. \quad (10)$$

and

$$P_f = P_2 (R_2/R_f)^{3\gamma}. \quad (11)$$

As the change in moles due to reactions in this case was found to be negligible ( $< 3\%$ ), the moles in the bubble are:

$$n_w = (4/3) \pi R_2^3 (P_S/R_g T). \quad (12)$$

Prasad Naidu *et al*, 1994 calculated various parameters like  $\gamma$ ,  $\sigma$  etc. for appropriate compositions and determined that the collapse temperatures fell between 2100 and 3200 K approximately, whereas the pressures varied between 80 and 120 12 MN/m<sup>2</sup>.

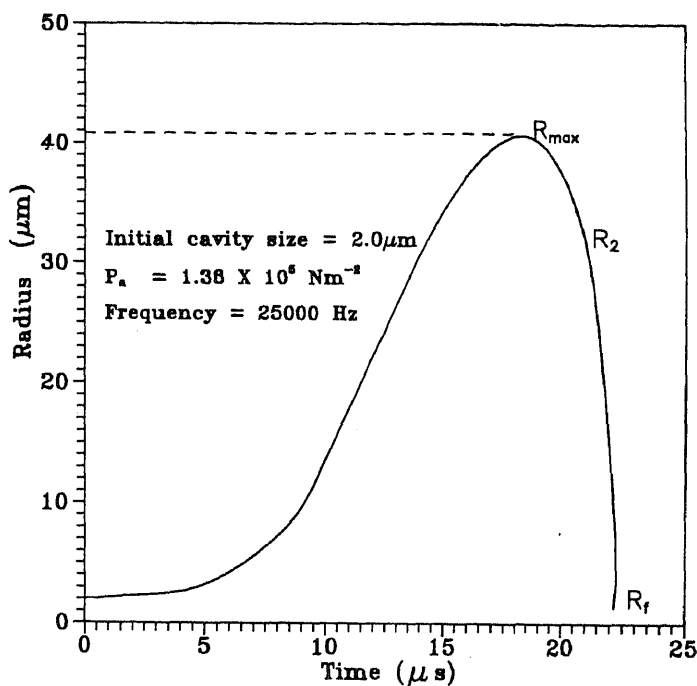
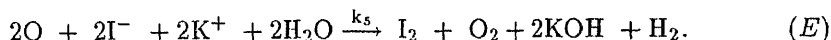
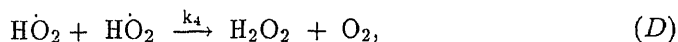
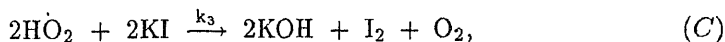
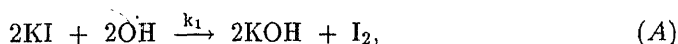


Figure 1. Curve showing typical radius versus time history.

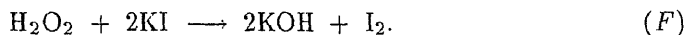
A typical profile of growth and collapse for one of their systems is shown in figure 1.

These temperature and pressure profiles were then used along with the bubble compositions (at  $R_2$ ) to evaluate the equilibrium compositions obtaining at  $R_f$ . This was done by calculating the composition leading to the minimization of free energy, through a program SOLGASMIX (Erickson 1975). The program yields as output the equilibrium compositions and fugacities of various species for any given temperature, pressure and input concentrations. The expected products have also to be provided, which in the case of water were  $H_2O$ , the constituents of the gas,  $O_2$ ,  $H$ ,  $OH$  and  $HO_2$ . The calculations showed significant formation of hydroxyl radicals ( $\sim 10^{-3}$  moles of  $OH$  per mole of water)

The various radicals provided by the SOLGASMIX program are released into the liquid phase, which is assumed to be well mixed. The reactions taking place in the liquid are given by Hart & Henglein (1985) and are



If sonication is carried out in the presence of the catalyst, ammonium molybdate,  $\text{H}_2\text{O}_2$  can also oxidise the iodide as



Based on the above, Prasad Naidu *et al* (1994) wrote the following material balance equations:

$$\text{Hydroxyl radicals:} \quad r_{\text{md}} C_A / dt = n \overline{C_A} - k_2 C_A^2 - k_1 C_A C_B \quad (13)$$

$$\text{Iodine:} \quad dC_C / dt = \beta_1 k_1 C_A C_B + \beta_2 k_4 C_B C_D + \beta_3 k_5 C_B C_E \quad (14)$$

$$\text{Hydroperoxyl radicals:} \quad dC_D / dt = n \overline{C_D} - k_3 C_D^2 - k_4 C_D C_B \quad (15)$$

$$\text{Oxygen atoms:} \quad dC_E / dt = n \overline{C_E} - k_5 C_E C_B \quad (16)$$

The rate at which the radicals are released into the liquid phase  $n \overline{C_i}$  reaches a steady state soon after the sonication begins. Thus radical concentration can reach a quasisteady state. If simultaneously  $C_B$  is sufficiently large that its concentration does not change significantly during the experiment, the rate of liberation of iodine also reaches a quasisteady state. Thus, the model can easily explain the key features of the findings of Hart & Henglein (1985). The model requires the number of collapsing cavities per unit time. The authors have found this from experiments conducted with ammonium molybdate where all OH go to the formation of iodine. The 'n' was found by them to be  $2.6445 \times 10^{10} / \text{l.s.}$  They found the values of rate constants by using the data with air as dissolved gas. Their predictions for oxygen atmosphere, along with their data are presented in figure 2. The agreement is quite good. The most interesting feature of this model is to be able to predict the maximum where oxygen-argon mixtures are employed. These predictions along with the data are presented in figure 3. The predictions are shown for two  $R_0$  values viz. 2.0 and  $2.5 \mu\text{m}$ . Though the rates are not predicted very well, the model does predict the maximum at around the same oxygen-air composition.

At best the model can be a reasonable approximation of the real phenomenon and captures its gross features. However, it neglects heat transport and cannot explain the influence of gas thermal conductivity on  $\text{H}_2\text{O}_2$  formation, which has been experimentally observed.

The number of nuclei have been evaluated from experiments. This may not be feasible for other reactions. The fate of the bubble is sensitive to the initial radius  $R_0$ . The model does not provide a foolproof method of assessing this. Even though, the collapse phase lasts for less than half a microsecond, the model assumes with success that thermodynamic equilibrium is attained. This assumption appears to be reasonable in view of the good experimental fit. However, more work on model systems, whose kinetics are known, needs to be done to have a more solid foundation for the assumption.

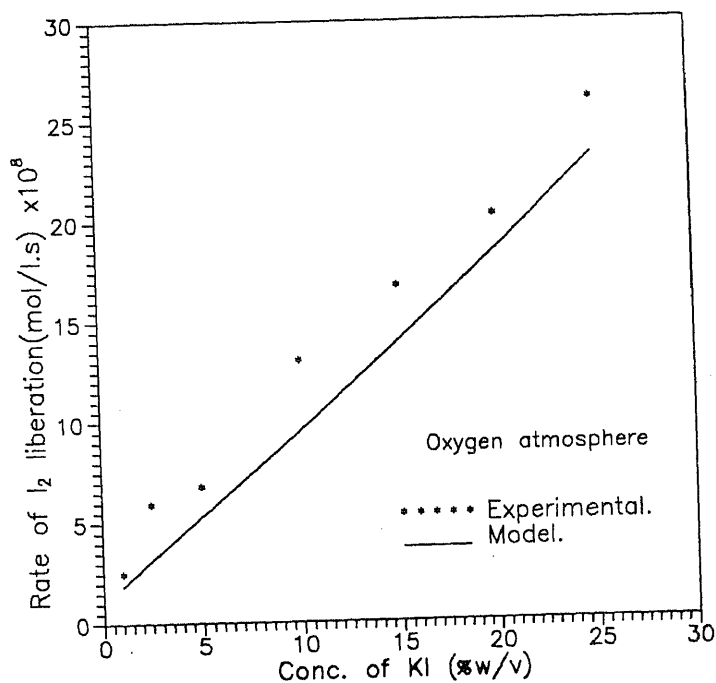


Figure 2. Comparison of model predictions with observation of iodine liberation rates: oxygen atmosphere.

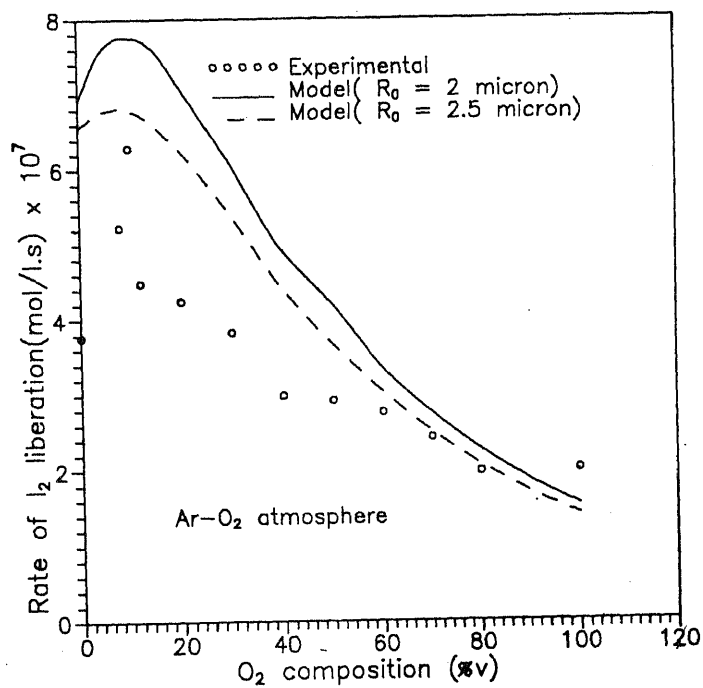


Figure 3. Comparison of model predictions with observations of iodine liberation rates under argon-oxygen atmosphere.

The weakest part of the model is the assumption about the state of mixing in the reactor. Though the assumption of complete mixing works reasonably well, the reactions are so fast that the time scale of mixing is likely to be much larger than the time scale of various reactions. Future work will possibly be based on more rigorous analysis.

No other homogeneous reaction has as yet been modelled.

## 6. Hypotheses about the effects of ultrasound on heterogeneous reactions

The situation with heterogeneous reactions, which are more numerous and important than homogeneous reactions, is very complex, and no models are available at present.

There are several effects of ultrasonication on heterogeneous reactions. The microjets created by the collapsing cavities cause strong convection. The convection can very often reduce diffusion limitations and thus can have an accelerating effect on any diffusion limited catalytic reactions. Perhaps, many effects of ultrasound on heterogeneous reactions fall under this broad class of explanation. The action of ultrasound on fragmentation of particles has been mentioned earlier. One obvious effect of this is the increased surface area in the system. Enhanced surface areas can have a definite effect on the rate of heterogeneous reactions. The pitting action of shock waves and microjets generated by collapsing bubbles on solid surfaces has already been mentioned. This could lead to an increased surface area, and perhaps a more active catalyst. Another possible influence of ultrasound is to alter the morphology of the solid surfaces exposed. For example, Suslick (1989) has explained the fusion of zinc particles to each other as due to shock waves. It is likely that the fused portion of the particles would have a different structure and hence its catalytic effect on a reaction is altered.

It is quite clear that a general model of the effect of ultrasound on heterogeneous reactions would be very difficult to develop at this stage. We report here some results on the modelling of a heterogeneous reaction based on the hypothesis that the convection caused by microjets is the sole effect of ultrasound.

## 7. Study of a heterogeneous reaction

The reaction between zinc and nickel chloride is an example of a galvanic reaction called cementation. The reaction is accompanied by evolution of hydrogen gas. In this particular case the cementation reaction involves the dissolution of the active metal (zinc) with the simultaneous deposition of the noble metal (nickel) present in solution in ionic form. The discrete nickel-plated zones on the particle surface then act as catalytic sites where electrochemical reduction of hydrogen ions present in the solution to neutral hydrogen atoms take place. The electrons needed for the reduction process are supplied by the dissolution of zinc, which goes into solution as zinc ions.

Analysis of products obtained from sonicated suspensions of zinc dust in aqueous solutions of nickel chloride showed (Jhansi 1992) that the above mechanism does apply. Further, experiments also showed (Jhansi 1992) that the rate of evolution of hydrogen was much faster from sonicated suspensions in comparison with stirred

suspensions. As mentioned earlier, reduction in mass transport resistance due to ultrasonication has been proposed as a possible cause for enhancement of reaction rates. We adopt this hypothesis here to explain the enhanced hydrogen evolution rates under the influence of ultrasound.

The following sequence of events can be proposed to explain the observations. When zinc is brought in contact with nickel chloride solution, nickel begins to plate on zinc. Simultaneously there is evolution of hydrogen. When nickel is in short supply, the nickel deposits are unable to cover the entire surface of zinc particles completely. Thus, the particle surface will have crevices or pits which enable the electrolyte to come in contact with the bare surface of zinc, where corrosion takes place ( $\text{Zn} \rightarrow \text{Zn}^{++} + 2e$ ) leading to deepening or growth of pits. The dynamics of the hydrogen evolution reaction is controlled by the rate of growth of pits. The growth of pits is usually diffusion-controlled. It was proposed (Banerjee *et al* 1994) that under these conditions pit growth is ohmically controlled under the influence of ultrasound,

As there is no current flow outside a particle in a cementation system, the following equation holds for a single particle :

$$A_{\text{Zn}} i_{\text{Zn/Zn}} + A_{\text{Ni}} i_{\text{H}_2/\text{Ni}} = 0 \quad (17)$$

and the equation for volumetric rate of hydrogen evolution can be written as

$$\dot{V}_{\text{H}_2}^{\text{particle}} = (R_g T / n_c F P) (A_{\text{Ni}} i_{\text{H}_2/\text{Ni}}). \quad (18)$$

### 7.1 Mathematical formulation

As the system is complex a number of idealizations have to be invoked before any quantitative analysis becomes feasible.

It will be assumed that the nickel coverage reaches quickly a steady value. It is assumed that there is no interaction between the growing pits on the surface of the zinc particle. The size and shape of the pits are also considered to be identical. Since the pit dimensions are very small ( $\sim 10^{-5} \text{ cm}$ ), the particle surface is idealized as a flat plane. It is assumed that metal dissolution takes place only from the pit bottom and the pit walls are passive. The dynamics of the pit growth will be modelled on the basis of quasi-steady state assumption. Thus if  $\Delta\phi_{\text{total}}$  is the total voltage drop available to drive the corrosion reaction :

$$\Delta\phi_{\text{total}} = \Delta\phi_{\text{ohmic}} + \Delta\phi_{\text{kinetic}}. \quad (19)$$

For typical parameter values corresponding to the conditions of experimentation, it can be shown (Banerjee 1994) that for a potential drop of 0.5V across the pit depth,  $\Delta\phi_{\text{kinetic}}$  is negligible even for a very small depth ( $\sim 10^{-4} \text{ cm}$ ).

The ohmic voltage drop consists of two parts, (i) the drop inside the pit and (ii) the drop outside. The voltage drop inside the pit can be obtained from straightforward application of the Ohm's law and is given as :

$$\Delta\phi_{\text{ohmic}}^{\text{in}} = (hi_{\text{Zn/Zn}}) / \kappa. \quad (20)$$

It can be shown that for pits of small diameter to depth ratios, as encountered in the present case, the voltage drop outside the pit can be neglected.

The total voltage drop is equal to the difference between the equilibrium potential values of the two electrodes corrected for concentration.

$$\Delta\phi_{total} = \phi_{cathode} - \phi_{anode}. \quad (21)$$

Hence for ohmically controlled situation where the dissolution current density is controlled solely by the voltage drop inside the pit we have the following governing equations for dissolution current density (Heimgartner & Bonhi 1985)

$$i_{Zn/Zn} = [n_a F \rho \Delta\phi_{total} \kappa / 2Mt]^{1/2}. \quad (22)$$

The hydrogen evolution rate can be written as :

$$\dot{V}_{H_2} = \frac{A_p N_p W_{Zn} R_g T}{n_c F P} \left[ \frac{n_a F \rho \Delta\phi_{total} \kappa}{2Mt} \right]^{1/2} = K \frac{1}{t^{1/2}}. \quad (23)$$

A similar analysis for conditions where diffusion controls the corrosion rates, the expression for current density from a single pit can be written as (Heimgartner & Bonhi 1985)

$$i_{Zn/Zn} = [n_a^2 F^2 \rho D (C_s - C_b) / 2Mt]^{1/2}, \quad (24)$$

and the total evolution rate of hydrogen is given by :

$$\dot{V}_{H_2} = \frac{A_p N_p W_{Zn} R_g T}{n_c F P} \left[ \frac{n_a^2 F^2 \rho D C_s}{2Mt} \right]^{1/2}, \quad (25)$$

since  $C_s \gg C_b$

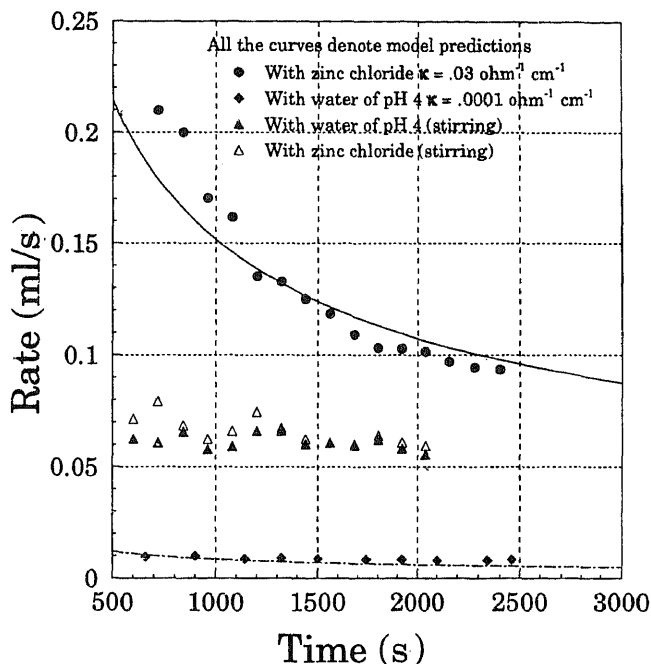
The above models contain several parameters. Of these,  $\Delta\phi_{total}$ ,  $\kappa$ ,  $C_s$  can be estimated *a priori*. Others e.g.  $N_{part}$ ,  $A_{par}$ ,  $tN_p$ ,  $A_p$ ,  $v$ ,  $N_0$  etc. can either be not evaluated or reliable data on these parameters does not exist in the literature. Thus, the models can be tested for predicting the trends of the data only after fixing the values of parameters from some experiments.

Experiments were conducted (Banerjee 1994) using an aqueous medium consisting of (2.4 g) zinc and (1 g ; 4.2 mmol) nickel chloride hexahydrate, and measurements of the rate of hydrogen evolution versus time were made under sonication. The samples were sonicated on an average for 30 minutes. The rate of hydrogen liberation was measured by using a soapfilm meter.

The results of the above experiment were used to fit the experimental data by treating the parameter  $K$  as adjustable, for the pitting model developed. The value of  $K_{ref}$  3.8 ml s<sup>-1/2</sup>.

It is possible from the above to predict the changes with variations in the experimental conditions. For example, if the value of conductivity is changed keeping all other conditions constant,  $K$  should change as the square root of the ratio of conductivities. Thus hydrogen evolution rates were measured with preformed nickel coated zinc dust suspended in acidulated water of pH 4 ( $\kappa \approx 0.0001$  ohm<sup>-1</sup>cm<sup>-1</sup>) and zinc chloride ( $\kappa \approx 0.03$  ohm<sup>-1</sup>cm<sup>-1</sup>). The hydrogen evolution rate was predicted and a comparison with the results of the experiments carried out with zinc chloride and acidulated water is shown in figure 4. The predictions are in good





**Figure 4.** Model predictions and observations for hydrogen liberation rates.

agreement with the experimental data.

From (25) it can be seen that the diffusion-controlled pit growth also yields a similar dependence of hydrogen evolution with time as the ohmically controlled one (see (23)). The diffusion-controlled model would not predict the effect of the change in ionic conductivity just described. An interesting reversal of the argument can be used to further confirm the predictions of the model. Under diffusion-controlled conditions, since electrical conductivity is not an important parameter, the model would predict that the rates of hydrogen evolution would be identical in aqueous solutions of zinc chloride as well as in acidulated water. The results of such an experiment are also shown in figure 4. Once again the model predictions are vindicated.

## 8. Large reactors and problems of scale up

Mason (1991) has given an excellent state of art review on the larger reactors as well as flow systems. Before discussing the problems associated with scale up, it is appropriate to briefly describe the larger systems which have been devised.

- (a) If only liquid-liquid systems are involved and we require large interfacial area, we can use liquid whistles, which are low intensity devices. These find use in industry for emulsification, but can be run as a flow system.
- (b) A variety of cleaning baths are available, which can be adapted to both the batch and flow systems. For example, a reaction vessel can be kept in a cleaning bath and can be operated in both batch and flow modes. Alternatively, external transducers can be attached to the walls of the reactors. There is

an advantage in this configuration that it permits the use of higher temperatures and pressures. The disadvantage of this system is that only low intensity ultrasound can be provided to the vessel contents.

One of the important considerations during the working of a large reactor is the region in which cavitations may occur. Good cavitation is normally observed up to a few centimeters from the wall. The second consideration is the damage to the reactor wall itself due to cavitation. Because of these reasons, the choice of the reactor configuration turns out to be a loop reactor, where a part of the liquid contents only is being sonicated at any particular time.

A sonicated section in a loop or flow reactor normally takes the form of a tube. The simplest of these is to have a standard pipe with ultrasonic probes inserted through T-joints at intervals. The main problems with the kind of sonicator is the tip erosion and uniform distribution of cavitation. In the Bronson sonochemical reactor, a coupling fluid is used in conjunction with the horns. The coupling fluid is chosen in such a way that it does not cavitate, but merely transmits energy to the wall.

One of the major problems associated with using pipes is that it is difficult to find transducers which make area contacts with the curvature of the pipe. To overcome this problem, both pentagonal and hexagonal pipes have been employed. Here focusing of energy can be obtained, without severe cavitation damage to the wall. However, the energy is focused near the centre line and less cavitation occurs near the walls. Thus the uniformity of cavitation cannot be ensured.

Two other kinds of tube reactors have been reported. In one case there are two coaxial tubes. The reaction mixture is passed through the inner tube whereas the coupling fluid (non cavitating) is passed through the outer one. In the second configuration the coolant can be passed through the inner tube and the reagent passed through the outer one. In each case the transducers are mounted in the outer tubes. Solids can also be handled by such reactors.

A loop reactor has been developed at Harwell, which consists of a stirred vessel, a heat exchanger and a sonicator. The reaction mixture is pumped from the vessel to the sonicator assembly through the heat exchanger. Two phase liquid-solid and liquid-liquid reactions have been conducted in the reactor.

### 8.1 *Scale up problems*

The first requirement for an engineering design is to be able to make an appropriate energy balance. The idea would be to calculate the percentage of the energy which goes towards cavitation. This has not been successfully achieved so far. In the absence of appropriate energy balance, use is normally made of the concept of sonochemical yield (SY) which is defined as the measured effect (say moles of a reactant reacted) per unit power input. Unfortunately, for the same power input per unit volume of the reactor, we do not get the same value of the sonochemical yield for two reactors of different sizes even if geometric similarity is maintained. The reason for this is the highly nonlinear nature of the phenomena associated with cavitation. Whereas reasonably uniform cavitation may be possible on the laboratory scale, the same cannot be achieved on a larger scale where the cavitation occurs only in a limited zone of the reactor.

At present, attempts are being made to identify different geometries and configurations which can give reasonably uniform acoustic intensity in the sonication zone. However, methods are not available by which the performance of a larger reactor can be predicted by doing experiments on a small batch reactor. Matters become more complex if a continuous reactor is to be designed based on the data collected on a batch system.

When heterogeneous reactions are involved, the particles undergo pitting and breakage during cavitation. Perhaps the particles could be sonicated independently of the reactor without the reactants and the resulting particles used. Then during sonication no significant breakage may result. However the increase in mass transport may not be the same for small and large reactors.

A further interesting feature has been introduced by Martin & Ward (1992), who bring out the "memory effect" in such reactors. When a catalyst or a solid reactant gets sonicated, its surface gets partially cleaned and its activity increases. But this activity does not fall to zero when the particle leaves the sonicator zone. It retains this activity for some time to come. These authors consider an exponential decay, but memory effects may be more complex and need to be investigated possibly through pulse sonolysis. It is possible that the memory effects last for a period lesser than the average residence time in the stirred vessel. To ensure that only spent particles go to the sonicated zone, it may be better to have a stirred sectionalised column which will act as a number of stirred vessels connected in series.

Again there are designs where the particles are packed. They will seriously scatter the ultra sound. The passage through the sonicator of fluidised particles offers a smaller alternation, but the sonicated zone may have to be enlarged.

Finally we have no control over the rate of nucleation. A small and a big reactor may have different rates of nucleation. It is necessary to study the nucleation phenomena and try to find external nucleation sites which can be introduced as well. Alternately a stage might be introduced in the loop where the reactants can be saturated with an inert gas. This will introduce some uniform nucleation rate with time as otherwise the contents slowly get degassed thereby affecting nucleation.

## 9. Conclusions

Sonochemistry as a discipline has grown significantly over the past decade. It provides conditions to conduct reactions which are not normally available to chemists with the existing techniques. Thus, through the use of power ultrasound, we can not only obtain dramatically higher rates, but also change the pathways of the reactions. Under ambient conditions we can conduct reactions which would otherwise require severe conditions. We can thus use less hazardous conditions and relatively impure reactants.

From a chemical engineering point of view we do not have examples where we can quantitatively predict the performance of such reactors, even in a batch system. However, we have reached a stage where a relatively gross model has been developed for sonolysis of KI solution in water. This could be used as a starting point for developing more robust models.

The industry has not yet started using such reactors, but one hopes that they will eventually find use, particularly for heterogeneous reactions involving liquids and solids, which at present have to be either purified or where severe conditions of

operation are involved. In the mean time, modelling and scale up of sonochemical reactors and development of new configurations provide a fertile area of research for chemical engineers.

## List of symbols

$A_{Zn}$	area of exposed zinc surface per particle, $cm^2$ ;
$A_{Ni}$	nickel plated area per particle, $cm^2$ ;
$A_p$	area of pit bottom, $cm^2$ ;
$C_b$	concentration in bulk solution, mole $cm^{-3}$ ;
$C_s$	concentration in the precipitated film, mole $cm^{-3}$ ;
$C_i$	concentration of $i$ th species, $gmol\ l^{-1}$ ;
$\bar{C}_i$	moles of $i$ th type of radicals/atoms in a single bubble, $gmol$ ;
$C_p$	specific heat at constant pressure, $cal\ gmol^{-1}\ ^\circ C^{-1}$ ;
$C_v$	specific heat at constant volume, $cal\ gmol^{-1}\ ^\circ C^{-1}$ ;
$D$	coefficient of diffusivity, $cm^2\ s^{-1}$ ;
$F$	Faraday's constant, $96487\ C\ equivalent^{-1}$ ;
$h$	depth of a pit, $cm$ ;
$i_{Zn/Zn}$	corrosion current density, $A\ cm^{-2}$ ;
$i_{H_2/Ni}$	current density for the hydrogen evolution reaction, $A\ cm^{-2}$ ;
$K$	pitting constant, $cm^3\ s^{-1/2}$ ;
$k_i$	reaction rate constants, $l\ gmol^{-1}\ s^{-1}$ , $i = 1, 2, 3, 4, 5$ ;
$M$	molecular weight, $g\ mole^{-1}$ ;
$n_c$	number of electrons exchanged (cathodic process);
$N_p$	total number of pits per gm of zinc metal, $gm^{-1}$ ;
$n$	number of bubbles collapsed per unit volume per unit time, $l^{-1}\ s^{-1}$ ;
$n_w$	number of moles of water vapor present in a single bubble at $R_2$ ;
$P$	pressure field, $N\ m^{-2}$ ;
$P_2$	pressure inside bubble when $R = R_2$ , $N\ m^{-2}$ ;
$P_{\infty}$	time varying pressure field, $N\ m^{-2}$ ;
$P_a$	pressure amplitude of ultrasonic field, $N\ m^{-2}$ ;
$P_b$	atmospheric pressure, $N\ m^{-2}$ ;
$P_f$	collapse pressure when $R = R_f$ , $N\ m^{-2}$ ;
$P_g$	partial pressure of the gas inside the bubble, $N\ m^{-2}$ ;
$P_{g0}$	initial gas pressure inside the bubble, $N\ m^{-2}$ ;
$P_{H_2O}$	vapour pressure of pure water at $298K$ , $N\ m^{-2}$ ;
$P_i$	total pressure inside the bubble, $N\ m^{-2}$ ;
$P_L$	pressure in the liquid just at the bubble surface, $N\ m^{-2}$ ;
$P_{LC}$	pressure corresponding to the radius of critical nucleus, $N\ m^{-2}$ ;
$P_s$	saturation vapour pressure of the liquid, $N\ m^{-2}$ ;
$P_v$	partial pressure of vapour in the bubble, $N\ m^{-2}$ ;
$R$	radius of the bubble at any time, $m$ ;
$R_0$	radius of the initial nucleus, $m$ ;
$R_2$	radius of the bubble at the beginning of the adiabatic collapse, $m$ ;
$R_c$	radius of critical nucleus, $m$ ;
$R_f$	collapse radius of the bubble, $m$ ;
$R_g$	universal gas constant $= 8.314\ J\ mol^{-1}\ K^{-1}$ ;
$\dot{R}$	surface velocity of the bubble, $m\ s^{-1}$ ;

$\ddot{R}$	acceleration of the surface of the bubble, $\text{m s}^{-2}$ ;
$T$	absolute temperature, K;
$T_2$	temperature inside bubble when $R = R_2$ , K;
$T_f$	collapse temperature when $R = R_f$ , K;
$t$	time, s;
$\alpha$	isotropic coefficient;
$\beta$	stoichiometric coefficient;
$\sigma$	surface tension of the liquid, $\text{N m}^{-1}$ ;
$\gamma$	ratio of specific heats;
$\rho_L$	density of the liquid, $\text{kg m}^{-3}$ ;
$\omega$	angular frequency, $\text{rad s}^{-1}$ ;
$\kappa$	conductivity of the solution, $\text{ohm}^{-1} \text{cm}^{-1}$ ;
$\phi$	electric potential, V;
$\phi_{\text{cathode}}$	cathodic potential, V;
$\phi_{\text{anode}}$	anodic potential, V;
$\Delta\phi_{\text{total}}$	total potential drop, V;
$\Delta\phi_{\text{ohmic}}^{\text{in}}$	potential drop inside the pit, V;
$\Delta\phi_{\text{ohmic}}^{\text{out}}$	potential drop outside the pit, V.

### Subscripts

A	OH;
B	KI;
C	I <sub>2</sub> ;
D	HO <sub>2</sub> ;
E	O.

## References

### General references

- Lorimer J P, Mason T J 1987 Sonochemistry, Part 1 - The physical aspects. *Chem. Soc. Rev.* 16: 239-274
- Mason T J (ed.) 1990 *Sonochemistry: The uses of ultrasound in chemistry* (London: Royal Society of Chemistry)
- Suslick K S (ed.) 1989 *Ultrasound. its chemical, physical, and biological effects* VCH

### Specific References

- Arakeri V H, Chakraborty S 1990 Studies towards potential use of ultrasonics in hydrodynamic cavitation control. *Curr. Sci.* 59: 1326-1333
- Boudjouk P, Thompson D P, Ohrbom W H, Han B H 1986 Effects of ultrasonic waves on the generation and reactivities of some metal powders. *Organometallics* 5: 1257-1260

- Eriksson G 1975 Thermodynamic studies of high temperature equilibria. XII SOL-GASMIX, a computer program for calculation of equilibrium composition in multiphase systems. *Chem. Scr.* 8: 100-103
- Flint E B, Suslick K S 1991 The temperature of cavitation. *Science* 246: 1397-1399
- Flynn H G 1964 Physics of acoustic cavitation in liquids. In *Physical acoustics* (ed.) W P Mason (New York: Academic Press) vol. 1B, pp 57-172
- Hart E J, Henglein A 1985 Free radical and free atom reactions in the sonolysis of aqueous iodide and formate solutions. *J. Phys. Chem.* 89: 4342-4347
- Heimgartner P, Bonhi H 1985 Mechanistic aspects of pit growth on nickel in diluted chloride solutions. *Corrosion* 41: 715
- Ley S V, Low C M R 1989 *Ultrasound in synthesis*. (New York: Springer-Verlag)
- Lush P A, Wood R J K, Carpanini L J 1983 *Proc. 6th Int. Conf. Erosion by Liquid and Solid Impact* (Cambridge: Cavendish Laboratory) paper 5
- Martin P D 1993 Sonochemistry in industry : Progress and prospects. *Chem. Ind.* 5 April: 233-236
- Martin P D, Ward L D 1992 Reactor design for sonochemical engineering. *Trans. I. Chem. E.* A70: 296-303
- Mason T J, Lorimer J P, Walton D J 1990 Sonochemistry. *Ultrasonics* 28: 333-337
- Mason T J 1991 *Practical sonochemistry: Users guide to applications in chemistry and chemical engineering* (Chichester: Ellis Horwood)
- Plesset M S 1949 The dynamics of cavitation bubbles. *Trans. ASME., J. Appl. Mech.* 71: 277-282
- Prasad Naidu D V, Rajan R, Kumar R, Gandhi K S, Arakeri V H, Chandrasekharan S 1995 Modelling of a batch sonochemical reactor. *Chem. Eng. Sci.* (in press)
- Price G J, Norris D J, West P J 1992 Polymerisation of methyl methacrylate initiated by ultrasound. *Macromolecules* 25: 6447-6454
- Repie O, Vogt S 1982 Ultrasound in organic synthesis: Cyclopropanation of olefins with zinc diiodomethane. *Tetrahedron Lett* 23: 2729-2732
- Sinisterra J V 1992 Application of ultrasound on biotechnology: An overview. *Ultrasonics* 30: 180-185
- Suslick K S 1989 The chemical effects of ultrasound. *Sci. Am.* February: 62-68
- Suslick K S 1990 Sonochemistry. *Science* 247: 1439-1445
- Suslick K S, Choe S B, Cichowias A A, Grinstaff M W 1991 Sonochemical synthesis of amorphous iron. *Nature (London)* 353: 414-41

## Subject Index

- ATM  
 Traffic engineering in ATM networks: Current trends and future issues 1005
- Aerodynamic interactions  
 Recent progress in dynamics and aeroelasticity 487
- Aerodynamical configurations  
 Detailed study of complex flow fields of aerodynamical configurations by using numerical methods 361
- Aeroelastic tailoring  
 Recent progress in dynamics and aeroelasticity 487
- Aeroelasticity  
 Recent progress in dynamics and aeroelasticity 487
- Aeroservoelasticity  
 Recent progress in dynamics and aeroelasticity 487
- Algorithms  
 Rudiments of complexity theory for scientists and engineers 985
- Annular  
 Interaction between a buoyancy-driven flow and an array of annular cavities 705
- Anusaraka  
*Paninian* framework and its application to *Anusaraka* 113
- Articulated rotor systems  
 Design and analysis trends of helicopter rotor systems 427
- Artificial intelligence  
 Understanding design: Artificial intelligence as an explanatory paradigm 5  
 Legal counselling system 75  
 A survey of Indian logic from the point of view of computer science 971
- Artificial neural network  
 Artificial neural networks for pattern recognition 189
- Attention-worthiness  
 Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Barrier  
 Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Bay of Bengal  
 Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories 567
- Bay of Bengal cyclones  
 On the prediction of storm surges 583
- Biological neural network  
 Artificial neural networks for pattern recognition 189
- Bluff body aerodynamics  
 On the boundary-layer control through momentum injection: Studies with applications 401
- Boltzman schemes  
 PVU and wave-particle splitting schemes for Euler equations of gas dynamics 1027
- Boundary-layer control  
 On the boundary-layer control through momentum injection: Studies with applications 401
- Brownian networks  
 Large dynamic scheduling in manufacturing systems using Brownian approximations 891
- Buoyancy  
 Interaction between a buoyancy-driven flow and an array of annular cavities 705
- Buoyant plane plumes  
 Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Cavitation  
 Sonochemical reaction engineering 1055  
 Interaction between a buoyancy-driven flow and an array of annular cavities 705
- Character synthesis  
 Script recognition 257
- Chemical reaction kinetics  
 Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials 817
- Coarticulation  
 Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Cognition  
 A survey of Indian logic from the point of view of computer science 971

- Communication  
 Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Complex flow fields  
 Detailed study of complex flow fields of aerodynamical configurations by using numerical methods 361
- Complexity theory  
 Rudiments of complexity theory for scientists and engineers 985
- Composite structures  
 Development of robust finite elements for general purpose structural analysis 289
- Computational complexity  
 Rudiments of complexity theory for scientists and engineers 985
- Computational heat transfer  
 Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces 723
- Content word  
 Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Control  
 Traffic engineering in ATM networks: Current trends and future issues 1005
- Convergence criteria  
 Solid finite elements through three decades 271
- Creativity  
 Understanding design: Artificial intelligence as an explanatory paradigm 5
- Cursive script  
 Script recognition 257
- Cyclone alert and warnings  
 Tropical cyclone hazards and warning and disaster mitigation systems in India 551
- Cyclone disaster (distress) mitigation  
 Tropical cyclone hazards and warning and disaster mitigation systems in India 551
- Database  
 Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Degeneracy  
 Distortion, degeneracy and rezoning in finite elements - A survey 311
- Dempster-Shafer theory  
 A prototype expert system for interpretation of remote sensing image data 93
- Design  
 Understanding design: Artificial intelligence as an explanatory paradigm 5
- Deterministic methods  
 Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories 567
- Diffuser flow  
 Finite element analysis of internal flows with heat transfer 785
- Disaster mitigation strategies  
 Earthquake hazard in Indian region 609
- Disaster reduction  
 Managing disasters precipitated by natural hazards 543
- Distortion  
 Distortion, degeneracy and rezoning in finite elements - A survey 311
- Distributed algorithms  
 Managing interprocessor delays in distributed recursive algorithms 995
- 3-D laminar flows  
 Finite element analysis of internal flows with heat transfer 785
- Domain objects  
 A prototype expert system for interpretation of remote sensing image data 93
- Drag reduction  
 On the boundary-layer control through momentum injection: Studies with applications 401
- Drought remedial measures  
 On the origin, incidence and impact of droughts over India and remedial measures for their mitigation 597
- Droughts  
 On the origin, incidence and impact of droughts over India and remedial measures for their mitigation 597
- Dynamic programming  
 A tutorial survey of reinforcement learning 891
- Dynamic scheduling  
 Large dynamic scheduling in manufacturing systems using Brownian approximations 851
- Dynamical characteristics  
 Dynamical characteristics of wave-excited channel flow 349
- Earthquakes  
 Earthquake hazard in Indian region 609
- El-Nino  
 On the origin, incidence and impact of droughts over India and remedial measures for their mitigation 597
- Element technology  
 Development of robust finite elements for general purpose structural analysis 289
- Enclosure  
 Natural convection of a radiating fluid in a square enclosure with perfectly conducting end walls 751
- Enthalpy formulation  
 A novel enthalpy formulation for multidimen-



- sional solidification and melting of a pure substance 833
- Eulerian velocity correction method
- Finite element analysis of internal flows with heat transfer 785
- Expert systems
  - A prototype expert system for interpretation of remote sensing image data 93
- Explanatory model of design
  - Understanding design: Artificial intelligence as an explanatory paradigm 5
- External loads
  - Computational aircraft dynamics and loads 467
- Extra-variational techniques
  - Solid finite elements through three decades 271
- Eye and wall cloud
  - Tropical cyclone hazards and warning and disaster mitigation systems in India 551
- FEPACS
  - Development of robust finite elements for general purpose structural analysis 289
- False colour composite
  - A prototype expert system for interpretation of remote sensing image data 93
- Fault-trees
  - Fault diagnosis of machines 23
- Faults
  - Earthquake hazard in Indian region 609
- Feature matrices
  - Script recognition 257
- Finite difference computations
  - Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials 817
- Finite element
  - Distortion, degeneracy and rezoning in finite elements - A survey 311
- Finite element method
  - Development of robust finite elements for general purpose structural analysis 289
  - Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems 765
  - Finite element analysis of internal flows with heat transfer 785
- Finite volume method
  - Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems 765
- Finite-difference method
  - A novel enthalpy formulation for multidimensional solidification and melting of a pure substance 833
- First order
  - Magneto-visco-elastic surface waves in stressed conducting media 337
- Flight loads
  - Computational aircraft dynamics and loads 467
- Flow through ducts
  - Finite element analysis of internal flows with heat transfer 785
- Flow transition
  - Flow transitions and pattern selection of the Rayleigh-Benard problem in rectangular enclosures 649
- Forced vibration
  - Forced vibration and low-velocity impact of laminated composite plates 509
- Forced wavy wall motion
  - Dynamical characteristics of wave-excited channel flow 349
- Formal methods
  - Design of software for safety critical systems 941
- Formants
  - Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Function word
  - Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Functional central limit theorem
  - Large dynamic scheduling in manufacturing systems using Brownian approximations 891
- Games
  - Planning in bridge with thematic actions 171
- General purpose packages
  - Development of robust finite elements for general purpose structural analysis 289
- Glass furnaces
  - Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces 723
- Grid optimization
  - Distortion, degeneracy and rezoning in finite elements - A survey 311
- Heat transfer
  - Interaction between a buoyancy-driven flow and an array of annular cavities 705
- Heat transfer augmentation
  - Finite element analysis of internal flows with heat transfer 785
- Heat-flux cylinder
  - Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Heavy traffic approximations
  - Large dynamic scheduling in manufacturing systems using Brownian approximations 891
- Helicopter rotor systems
  - Design and analysis trends of helicopter rotor

- systems 427
- Higher-order networks
  - Convergence of higher-order two-state neural networks with modified updating 239
- Homeostasis
  - A fuzzy expert system approach using multiple experts for dynamic follow-up of endemic diseases 51
- Hopfield networks
  - Convergence of higher-order two-state neural networks with modified updating 239
- Hurricane intensity
  - Tropical cyclone hazards and warning and disaster mitigation systems in India 551
- Hydrostatic tension or compression
  - Magneto-visco-elastic surface waves in stressed conducting media 337
- Idiotypic network
  - A fuzzy expert system approach using multiple experts for dynamic follow-up of endemic diseases 51
- Image interpretation
  - A prototype expert system for interpretation of remote sensing image data 93
- Immunological reaction
  - A fuzzy expert system approach using multiple experts for dynamic follow-up of endemic diseases 51
- Indian logic
  - A survey of Indian logic from the point of view of computer science 971
- Initial stress
  - Magneto-visco-elastic surface waves in stressed conducting media 337
- Inlet velocity distortion
  - Finite element analysis of internal flows with heat transfer 785
- Internal flows
  - Finite element analysis of internal flows with heat transfer 785
- Interpolation functions
  - Solid finite elements through three decades 271
- Interpretation keys
  - A prototype expert system for interpretation of remote sensing image data 93
- Intonation
  - Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Invention
  - Understanding design: Artificial intelligence as an explanatory paradigm 5
- Jacobian
  - Distortion, degeneracy and rezoning in finite elements - A survey 311
- Karaka
  - Paninian* framework and its application to *Anusaraka* 113
- Knowledge
  - Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Knowledge structures
  - Planning in bridge with thematic actions 171
- Knowledge-based planning
  - Planning in bridge with thematic actions 171
- Laminated composites
  - Forced vibration and low-velocity impact of laminated composite plates 509
- Language
  - Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
  - A survey of Indian logic from the point of view of computer science 971
- Language accessor
  - Paninian* framework and its application to *Anusaraka* 113
- Layerwise theory
  - Forced vibration and low-velocity impact of laminated composite plates 509
- Legal expert systems
  - Legal counselling system 75
- Legal predictive system
  - Legal counselling system 75
- Legal reasoning
  - Legal counselling system 75
- Line heat source
  - Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Logic
  - A survey of Indian logic from the point of view of computer science 971
- Low-velocity impact
  - Forced vibration and low-velocity impact of laminated composite plates 509
- Machine translation
  - Paninian* framework and its application to *Anusaraka* 113
- Magneto-visco-elastic
  - Magneto-visco-elastic surface waves in stressed conducting media 337
- Magnitude
  - Earthquake hazard in Indian region 609
- Manoeuvre loads
  - Computational aircraft dynamics and loads 467
- Manufacturing systems
  - Large dynamic scheduling in manufacturing systems using Brownian approximations 891

- Mapping  
Distortion, degeneracy and rezoning in finite elements - A survey 311
- Mass media  
Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Model-based diagnosis  
Fault diagnosis of machines 23
- Modified updating  
Convergence of higher-order two-state neural networks with modified updating 239
- Multi-level qualitative reasoning  
Fault diagnosis of machines 23
- Multiclass networks  
Large dynamic scheduling in manufacturing systems using Brownian approximations 891
- Multidimensional phase-change problems  
A novel enthalpy formulation for multidimensional solidification and melting of a pure substance 833
- Multidisciplinary computational methods  
Computational aircraft dynamics and loads 467
- NO<sub>x</sub> formation  
Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces 723
- Natural convection  
The differentially heated cavity 619  
Natural convection of a radiating fluid in a square enclosure with perfectly conducting end walls 751
- Natural hazards  
Managing disasters precipitated by natural hazards 543
- Navier equation  
Solid finite elements through three decades 271
- Neural dynamics  
Convergence of higher-order two-state neural networks with modified updating 239
- Neural networks  
A tutorial survey of reinforcement learning 851
- Non-Boussinesq  
The differentially heated cavity 619
- Non-Newtonian materials  
Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials 817
- Nonlinear integral equation  
Forced vibration and low-velocity impact of laminated composite plates 509
- Numerical aerodynamics  
Detailed study of complex flow fields of aerodynamical configurations by using numerical methods 361
- Numerical plume simulations  
Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Numerical weather prediction  
Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories 567
- On-line and off-line recognition  
Script recognition 257
- Optimal control  
A tutorial survey of reinforcement learning 851
- Overlapping segments  
Script recognition 257
- Paninian* grammar  
*Paninian* framework and its application to *Anusaraka* 113
- Papcovich-Neuber solution  
Solid finite elements through three decades 271
- Pattern recognition  
Artificial neural networks for pattern recognition 189
- Pattern selection  
Flow transitions and pattern selection of the Rayleigh-Benard problem in rectangular enclosures 649
- Peculiar velocity based upwinding  
PVU and wave-particle splitting schemes for Euler equations of gas dynamics 1027
- Power system diagnosis  
Fault diagnosis of machines 23
- Prosodic features  
Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Qualitative model-based diagnosis  
Fault diagnosis of machines 23
- Quality of surveys  
Traffic engineering in ATM networks: Current trends and future issues 1005
- Quantisation  
Script recognition 257
- Radiation  
Natural convection of a radiating fluid in a square enclosure with perfectly conducting end walls 751
- Radiative heat transfer  
Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces 723
- Rayleigh-Benard problem  
Flow transitions and pattern selection of the Rayleigh-Benard problem in rectangular

- enclosures 649
- Rectangular enclosure
  - Flow transitions and pattern selection of the Rayleigh-Benard problem in rectangular enclosures 649
- Reinforcement learning
  - A tutorial survey of reinforcement learning 851
- Remote sensing
  - A prototype expert system for interpretation of remote sensing image data 93
- Rezoning
  - Distortion, degeneracy and rezoning in finite elements - A survey 311
- Rotorcraft
  - Design and analysis trends of helicopter rotor systems 427
- Route to chaos
  - Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Rule-based diagnosis
  - Fault diagnosis of machines 23
- Safety critical systems
  - Design of software for safety critical systems 941
- Script recognition
  - Script recognition 257
- Seismic hazard analysis
  - Earthquake hazard in Indian region 609
- Seismic source zones
  - Earthquake hazard in Indian region 609
- Seismicity data
  - Earthquake hazard in Indian region 609
- Shape vectors
  - Script recognition 257
- Single-screw extruder
  - Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials 817
- Software design
  - Design of software for safety critical systems 941
- Solid finite elements
  - Solid finite elements through three decades 271
- Sonochemistry
  - Sonochemical reaction engineering 1055
- Sonolysis
  - Sonochemical reaction engineering 1055
- Sonoreactors
  - Sonochemical reaction engineering 1055
- Spectral collocation method
  - Dynamical characteristics of wave-excited channel flow 349
- Stability characteristics
  - Dynamical characteristics of wave-excited channel flow 349
- Stochastic approximations
  - Managing interprocessor delays in distributed recursive algorithms 995
- Storm surges
  - On the prediction of storm surges 583
- Storm tracks
  - Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories 567
- Strict Liapunov systems
  - Managing interprocessor delays in distributed recursive algorithms 995
- Structural analysis
  - Development of robust finite elements for general purpose structural analysis 289
- Structural optimisation
  - Recent progress in dynamics and aeroelasticity 487
- Structural optimization
  - Design and analysis trends of helicopter rotor systems 427
- Surface waves
  - Magneto-visco-elastic surface waves in stressed conducting media 337
- Surge prediction techniques
  - On the prediction of storm surges 583
- Synchronous programming paradigm
  - Design of software for safety critical systems 941
- T-number
  - Tropical cyclone hazards and warning and disaster mitigation systems in India 551
- Tapering stepwise
  - Managing interprocessor delays in distributed recursive algorithms 995
- Temporal analysis
  - Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems 765
- Text-to-speech system
  - Significance of knowledge sources for a text-to-speech system for Indian languages 147
- Tract segments
  - Script recognition 257
- Traffic characterization
  - Traffic engineering in ATM networks: Current trends and future issues 1005
- Transient thermal analysis
  - Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems 765
- Transition
  - The differentially heated cavity 619
- Transport phenomena
  - Computational study of transport processes in a single-screw extruder for non-Newtonian chemi-

- cally reactive materials 817
- Tropical cyclones
  - Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories 567
- Turbulence
  - The differentially heated cavity 619
- Turbulence models
  - Finite element analysis of internal flows with heat transfer 785
- Turbulent combustion
  - Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces 723
- Ultrasound
  - Sonochemical reaction engineering 1055
- Uncertain environment
  - Planning in bridge with thematic actions 171
- Uncertainty handling
  - A prototype expert system for interpretation of remote sensing image data 93
- Upwind methods for Euler equations
  - PVU and wave-particle splitting schemes for Euler equations of gas dynamics 1027
- Wave-excited channel flow
  - Dynamical characteristics of wave-excited channel flow 349
- Wave-particle splitting
  - PVU and wave-particle splitting schemes for Euler equations of gas dynamics 1027
- Weak convergence
  - Large dynamic scheduling in manufacturing systems using Brownian approximations 891

## Author Index

- Acharya D P  
     *see* Das Samar Chandra 337
- Acharya S  
     Natural convection of a radiating fluid in a square enclosure with perfectly conducting end walls 751
- Appa K  
     Computational aircraft dynamics and loads 467
- Argyris J  
     *see* Appa K 467
- Arun Kumar A T  
     *see* Mahabala H N 23
- Aswathanarayana P A  
     Finite element analysis of internal flows with heat transfer 785
- Balakrishnan N  
     *see* Deshpande S M 1027
- Banerjee Apurba  
     A fuzzy expert system approach using multiple experts for dynamic follow-up of endemic diseases 51
- Basu Anupam  
     *see* Banerjee Apurba 51
- Belegundu Ashok D  
     *see* Salagame Raviprakash R 311
- Bharati Akshar  
     *Paninian* framework and its application to *Anusaraka* 113
- Borkar V S  
     Managing interprocessor delays in distributed recursive algorithms 995
- Chaitanya Vineet  
     *see* Bharati Akshar 113
- Chaudhary Shresh  
     Reducing barriers of communication across Indian languages: An AI and ES approach to mass media 129
- Chopra Inderjit  
     Design and analysis trends of helicopter rotor systems 427
- Das A  
     Detailed study of complex flow fields of aerodynamical configurations by using numerical methods 361
- Das P K  
     On the prediction of storm surges 583
- Das Samar Chandra  
     Magneto-visco-elastic surface waves in stressed conducting media 337
- Dasgupta Subrata  
     Understanding design: Artificial intelligence as an explanatory paradigm 5
- Date A W  
     A novel enthalpy formulation for multidimensional solidification and melting of a pure substance 833
- Deshpande S M  
     PVU and wave particle splitting schemes for Euler equations of gas dynamics 1027
- Desrayaud Gilles  
     Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Faghri M  
     Interaction between buoyancy-driven flow and an array of annular cavities 705
- Gandhi K S  
     Sonochemical reaction engineering 1055
- Gaur V K  
     Managing disasters precipitated by natural hazards 543
- Gopalakrishna S  
     Computational study of transport processes in a single-screw extruder for non-Newtonian chemically reactive materials 817
- Gravey A  
     Traffic engineering in ATM networks: Current trends and future issues 1005
- Hebuterne G  
     *see* Gravey A 1005
- Hoogendoorn C J  
     Computational modelling of turbulent flow, combustion and transfer in glass furnaces 723
- Jaluria Y  
     Foreword 615  
     Computational study of transport processes in a

- single-screw extruder for non-Newtonian chemically reactive materials 817
- Joseph P V  
Tropical cyclone hazards and warning and disaster mitigation systems in India 551
- Kapania Rakesh K  
see Nosier Asghar 509
- Khatti K N  
Earthquake hazard in Indian region 609
- Khemani Deepak  
Planning in bridge with thematic actions 171
- Koster C L  
Computational modelling of turbulent flow, combustion and heat transfer in glass furnaces 723
- Kumar R  
see Gandhi K S 1055
- Kurup R R  
see Mahabala H N 23
- Lakshminath A  
see Shashi M 75
- Lauriat Guy  
Buoyant plane plumes from heated horizontal confined wires and cylinders 671
- Madhukumar A S  
see Yegnanarayana B 147
- Mahabala H N  
Fault diagnosis of machines 23
- Majumder Arun Kumar  
see Banerjee Apurba 51
- Mazumdar R R  
see Gravey A 1005
- Modi V J  
On the boundary-layer control through momentum injection: Studies with applications 401
- Mohan Ram V  
Finite element/finite volume approaches with adaptive time stepping strategies for transient thermal problems 765
- Mohanty U C  
Tropical cyclones in the Bay of Bengal and deterministic methods for prediction of their trajectories 567
- Molki M  
Interaction between buoyancy-driven flow and an array of annular cavities 705
- Mooley D A  
Origin, incidence and impact of droughts over India and remedial measures for their mitigation 597
- Mukutmoni D  
Flow transitions and pattern selection of the Rayleigh-Benard problem in rectangular enclosures 649
- Naganarayana B P  
see Prathap G 289
- Narahari Y  
see Ravikumar K 891
- Nosier Asghar  
Forced vibration and low-velocity impact of laminated composite plates 509
- Panda Keshab  
see Upadhyaya A R 487
- Phansalkar V V  
see Borkar V S 995
- Paolucci S  
The differentially heated cavity 619
- Prathap G  
Development of robust finite elements for general purpose structural analysis 289
- Raghurama Rao S V  
see Deshpande S M 1027
- Rajendran S  
see Yegnanarayana B 147
- Raju K V S V N  
see Shashi M 75
- Ramachandran V R  
see Yegnanarayana B 147
- Rao P V S  
Script recognition 257
- Ravi Prakash G  
see Mahabala H N 23
- Ravikumar K  
Large dynamic scheduling in manufacturing systems using Brownian approximations 891
- Ravindran B  
see Sathya Keerthi S 851
- Ravisankar M S  
Finite element analysis of internal flows with heat transfer 785
- Reddy J N  
see Nosier Asghar 509
- Rosenberg C  
see Gravey A 1005
- Salagame Raviprakash R  
Distortion, degeneracy and rezoning in finite elements - A survey 311
- Sangal Rajeev  
see Bharati Akshar 113
- Sarma Chandra Sekhara L  
A prototype expert system for interpretation of remote sensing image data 93
- Sarma V V S  
A survey of Indian logic from the point of view of computer science 971  
see Chandra Sekhara Sarma L 93
- Sathya Keerthi S  
A tutorial survey of reinforcement learning 851

- Seetharamu K N  
Finite element analysis of internal flows with  
heat transfer 785
- Selvarajan S  
Dynamical characteristics of wave-excited chan-  
nel flow 349
- Sengupta P R  
see Das Samar Chandra 337
- Shashi M  
Legal counselling system 75
- Shrinivasa U  
see Venkatesh D N 271
- Shyamasundar R K  
Design of software critical systems algorithms  
941
- Somashekar B R  
see Prathip G 289
- Srinivas M  
Finite element analysis of internal flows with  
heat transfer 785
- Srinivasan J  
Foreword 615
- Tamma Kumar K  
Finite element/finite volume approaches with  
adaptive time stepping strategies for transient  
thermal problems 765
- Upadhy A R  
Recent progress in dynamics and aeroelasticity  
487
- Vinay V  
Rudiments of complexity theory for scientists  
and engineers 985
- Vasanta Ram V  
see Selvarajan S 349
- Venkatesh D N  
Solid finite elements through three decades 271
- Vidyasagar M  
Convergence of higher-order two-state neural net-  
works with modified updating 239
- Wieringa J A  
Computational modelling of turbulent flow, com-  
bustion and heat transfer in glass furnaces 723
- Williams M L  
Natural convection of a radiating fluid in a square  
enclosure with perfectly conducting end walls  
751
- Yang K T  
see Mukutmoni D 649
- Yegnanarayana B  
Significance of knowledge sources for a text-to-  
speech system for Indian languages 147  
Artificial neural networks for pattern recognition  
189
- Yokomizo T  
see Modi V J 401
- Yucel A  
Natural convection of a radiating fluid in a square  
enclosure with perfectly conducting end walls  
751



